xAILAB
EXPLAINABLE MACHINE LEARNING

Generative Data Augmentation in the Embedding Space of Vision Foundation Models to Address Long-Tailed Learning and Privacy Constraints

Master Thesis

Master of Science in Computing in the Humanities

David Tafler

June 12, 2024

Supervisor:

1st: Prof. Dr. Christian Ledig
2nd: Francesco Di Salvo

Chair of Explainable Machine Learning
Faculty of Information Systems and Applied Computer Sciences
Otto-Friedrich-University Bamberg

Abstract

This thesis explores the potential of generative data augmentation in the embedding space of vision foundation models, aiming to address the challenges of long-tailed learning and privacy constraints. Our work leverages Conditional Variational Autoencoders (CVAEs) to enrich the representation space for underrepresented classes in highly imbalanced datasets and to enhance data privacy without compromising utility. We develop and assess methods that generate synthetic data embeddings conditioned on class labels, which both mimic the distribution of original data for privacy purposes and augment data for tail classes to balance datasets. Our methodology shows that embedding-based augmentation can effectively improve classification accuracy in long-tailed scenarios by increasing the diversity and volume of minor class samples. Additionally, we demonstrate that our approach can generate data that maintains privacy through effective anonymization of embeddings. The outcomes suggest that generative augmentation in embedding spaces of foundation models offers a promising avenue for enhancing model robustness and data security in practical applications. The findings have significant implications for deploying machine learning models in sensitive domains, where data imbalance and privacy are critical concerns.

Abstract

Diese Arbeit untersucht das Potenzial der generativen Datenaugmentation im Embedding-Raum von Vision-Foundation-Modellen. Ziel ist es, die Herausforderungen des Long-Tailed Learning und von Datenschutzbeschränkungen anzugehen. Unsere Arbeit nutzt Conditional Variational Autoencoders (CVAEs), um den Embedding-Raum für unterrepräsentierte Klassen in stark unausgewogenen Datensätzen zu erweitern. Gleichzeitig soll die Datensicherheit erhöht werden, ohne die Nützlichkeit der Daten zu beeinträchtigen. Wir entwickeln und bewerten Methoden zur auf Klassenlabels konditionierten Erzeugung synthetischer Embeddings. Diese imitieren die Verteilung der Originaldaten zur Datenanonymisierung und augmentieren Minderheitsklassen, um Datensätze auszugleichen. Unsere Methodik zeigt, dass die auf Embeddings basierende Augmentation die Klassifikationsgenauigkeit in Long-Tailed-Szenarien effektiv verbessern kann, indem die Vielfalt und das Volumen in Minderheitsklassen erhöht werden. Zusätzlich demonstrieren wir, dass unser Ansatz Daten generieren kann, die durch effektive Anonymisierung von Embeddings die Privatsphäre wahren. Die Ergebnisse legen nahe, dass generative Augmentation im Embedding-Raum von Foundation-Modellen eine vielversprechende Möglichkeit bietet, die Robustheit von Modellen und die Datensicherheit in praktischen Anwendungen zu verbessern. Die Ergebnisse haben bedeutende Implikationen für den Einsatz von maschinellen Lernmodellen in sensiblen Bereichen, in denen Datenunausgewogenheit und Datenschutz von entscheidender Bedeutung sind.

Contents

List of Figures

List of Tables

List of Acronyms

ADASYN	Adaptive Synthetic Sampling
AI	Artificial Intelligence
CLIP	Contrastive Language-Image Pretraining
CNN	Convolutional Neural Network
CVAE	Conditional Variational Autoencoder
DINO	Self-Distillation with No labels
FID	Fréchet Inception Distance
GAN	Generative Adversarial Network
GPT	Generative Pre-trained Transformer
MLP	Multi-Layer Perceptron
NLP	Natural Language Processing
RLHF	Reinforcement Learning from Human Feedback
RNN	Recurrent Neural Network
SMOTE	Synthetic Minority Over-sampling Technique
VAE	Variational Autoencoder
ViT	Vision Transformer

Notation

Numbers, Arrays, and Linear Algebra

b	A scalar (integer or real)
$\boldsymbol{b}$	A vector
$\boldsymbol{B}$	A matrix
$\mathbf{I}$	The identity matrix with dimensionality implied by context
$\mathbf{0}$	The zero vector with dimensionality implied by context
$\|\|\boldsymbol{b}\|\|$	The L^2 norm of $\boldsymbol{b}$
$\mathrm{Tr}(\boldsymbol{B})$	The trace of $\boldsymbol{B}$

Datasets

$\mathcal{X}$	A set of training examples in the input space
$\mathcal{Z}$	A set of training examples in the feature space of an encoder
$\mathcal{A}$	A set of generated examples in the feature space of an encoder
x_i	The i-th example from a dataset in the input space
z_i	The i-th example from a dataset in the feature space of an encoder
l_i	The vector resulting from the mapping of z_i to the latent space of a CVAE
a_i	The i-th example from a generated dataset in the feature space of an encoder
y_i	The target associated with x_i, z_i, or a_i for supervised learning
$\mathcal{A} \cup \mathcal{Z}$	The union of sets $\mathcal{A}$ and $\mathcal{Z}$
ρ	The imbalance ratio of a dataset
C	The number of classes in a dataset
n_c	The number of samples in class c

Functions

$f(x; \theta)$	A function of x parametrized by θ. (Sometimes we write $f(x)$ and omit the argument θ to lighten notation)
$\log x$	Natural logarithm of x
$MSE(x, x^*)$	Mean Squared Error of estimator x^*

Probability

$P(\mathrm{a})$	A probability distribution over a discrete variable
$p(\mathrm{a})$ or $q(\mathrm{a})$	A probability distribution over a continuous variable, or over a variable whose type has not been specified
$\mathrm{a} \sim P$	Random variable a has distribution P
$\mathcal{N}(x; \mu, \Sigma)$	Gaussian distribution over x with mean μ and covariance Σ
$\mathcal{N}(\mu, \Sigma)$	Gaussian distribution with mean μ and covariance Σ
$D_{\mathrm{KL}}(P\|Q)$	Kullback-Leibler divergence of P and Q
$\mathbb{E}_{\mathrm{x}\sim P}[f(x)]$	Expectation of $f(x)$ with respect to $P(\mathrm{x})$

1 Introduction

Categorization is fundamental to human experience and the categories we learn and encounter throughout our lives follow certain statistical patterns. One such pattern is long-tailed imbalance: few categories have a very high frequency, while the majority of categories only occur rarely. For example, word frequencies in natural language are inversely proportional to their rank, approximating a power law known as Zipf's law (Zipf, 1935; Piantadosi, 2014). Categories of visual objects and object subcategories (Salakhutdinov et al., 2011; Zhu et al., 2014), and therefore many real-world datasets follow a similar long-tailed distribution (Van Horn et al., 2018; Guo et al., 2016; Zhang et al., 2017; Ju et al., 2021; Galdran et al., 2021; Zhang et al., 2023). Nevertheless, many common image classification datasets are artificially class-balanced (Deng et al., 2009b; Krizhevsky and Hinton, 2009; Lecun et al., 1998), and conventional empirical risk minimization algorithms perform poorly when trained with imbalanced data (Van Horn and Perona, 2017; He and Garcia, 2009; Buda et al., 2018; Zhang et al., 2023).

Among the most common methods to improve long-tailed classification are sampling strategies (Chawla et al., 2002; He et al., 2008; Kang et al., 2019) and the use of specially designed loss functions (Lin et al., 2017; Cao et al., 2019; Cui et al., 2019; Ren et al., 2020; Tan et al., 2020). While these methods often improve upon natural, instance-balanced sampling and a conventional cross-entropy loss function, they focus solely on re-sampling and re-weighting, not on increasing diversity in tail classes, which are the underrepresented categories in a dataset with a skewed class distribution. One usually increases diversity of training samples with data augmentation techniques and several such techniques have been developed for long-tailed learning. However, they often either require the training of a feature encoder (Yin et al., 2019; Liu et al., 2020; Wang et al., 2021; Chu et al., 2020), augment the data at the input-level (Kim et al., 2020; Fajardo et al., 2021; Dablain et al., 2022), or depend on the presence of a validation set (Zang et al., 2021; Li et al., 2021).

In this work, we focus on the setting of a frozen feature encoder and the absence of a validation set, for the following reasons. Firstly, while some artificially created long-tailed datasets include balanced validation sets (Liu et al., 2019b), real-world applications with long-tailed datasets often do not provide this luxury, due to the inherent difficulty in obtaining sufficient samples from underrepresented classes. Secondly, we use frozen feature encoders, among other reasons because of a growing interest in and availability of large-scale pretrained vision encoders. The embeddings of these foundation models provide a basis for many down-stream tasks, often without the need for fine-tuning (Bommasani et al., 2022; Caron et al., 2021; Oquab et al., 2023; Radford et al., 2021). Usually based on the transformer architecture (Vaswani et al., 2017; Dosovitskiy et al., 2020), vision foundation models inherit their robustness to distribution shift, corruptions, and natural adversarial examples (Paul and Chen, 2022). Furthermore, the number of domain-specific foundation models has been increasing in recent years (Li et al.; Mai et al., 2022; Nguyen et al., 2023; Tu et al., 2024; Zhou et al., 2023). This makes vision foundation models a

natural choice when approaching long-tailed recognition by decoupling the processes of representation learning and classifier training (Kang et al., 2019).

To address long-tailed classification in this setup, we train a Variational Autoencoder (VAE) on the embeddings (i.e., encoded features) of a long-tailed dataset, conditioning it on the class labels. This allows us to generate embeddings of arbitrary classes by sampling from the autoencoder's latent space. By combining generated embeddings with the embeddings of the long-tailed dataset, we aim to increase the meaningful diversity of tail classes in the feature space of a given vision foundation model. While traditional interpolation-based techniques can also be applied in this setup (Chawla et al., 2002; He et al., 2008; Chou et al., 2020), they presuppose specific assumptions about the semantics of interpolation in high-dimensional embedding spaces. For instance, one common assumption is that the linear interpolation between two embeddings of the same class results in an embedding that is also a member of this class. But while the true meaning of interpolation may be hidden in the feature encoder's weights, it is completely obscure to us. Therefore, this assumption might be wrong in many cases. In contrast, we *train* a model to produce samples using specific inductive biases by conditioning it on class labels and encouraging a normal distribution of samples in the its latent space.

The proposed method of training a Conditional Variational Autoencoder (CVAE) on the frozen embeddings of a given dataset is very general and can therefore be applied to problems beyond long-tailed learning. One such challenge and a second focus of this work lies in the domain of privacy and is concerned with anonymizing data while preserving its usefulness (Sweeney, 2002; Newton et al., 2005; Meden et al., 2018). This is particularly important when *sharing* sensitive data, such as medical records or surveillance footage. Instead of sharing the raw data itself, a naive solution might be to share the embeddings of a dataset (produced by encoding it with a given feature encoder), or to only share a model trained on the data, such as a classifier. However, attacks such as model inversion (Fredrikson et al., 2014, 2015) could still lead to the disclosure of the original data. Therefore, we propose to first locally capture the distribution of the original data with a CVAE. In a second step, only the frozen CVAE's decoder needs to be shared, allowing the recipient to generate samples according to the original distribution, without disclosing the dataset. This not only enhances privacy, but also drastically decreases the required amount of data to be exchanged. Figure 1 illustrates our proposed approach and its applications in both long-tailed learning and privacy preservation.

The remainder of this work is structured as follows. Section 2 introduces the topics of foundation models, long-tailed learning, and privacy as they relate to the generative data augmentation approach, and discusses relevant literature. In Section 3, we present our approach, introducing the relevant notation and detailing how we use generative models to rebalance long-tailed datasets and anonymize training data. Section 4 describes the setup and results of the experiments conducted to evaluate our approach. This includes an exploratory analysis of the latent-space properties of CVAEs, a quantitative evaluation of the quality of our generated samples, and their potential to increase diversity in classes with few examples. Notably, this section also

reports on experiments comparing the performance of our approach on long-tailed classification with commonly used state-of-the-art methods and evaluations of our anonymization technique using real-world medical datasets. In Section 5, we discuss the broader impact and limitations of our work. Finally, Section 6 provides a brief summary.

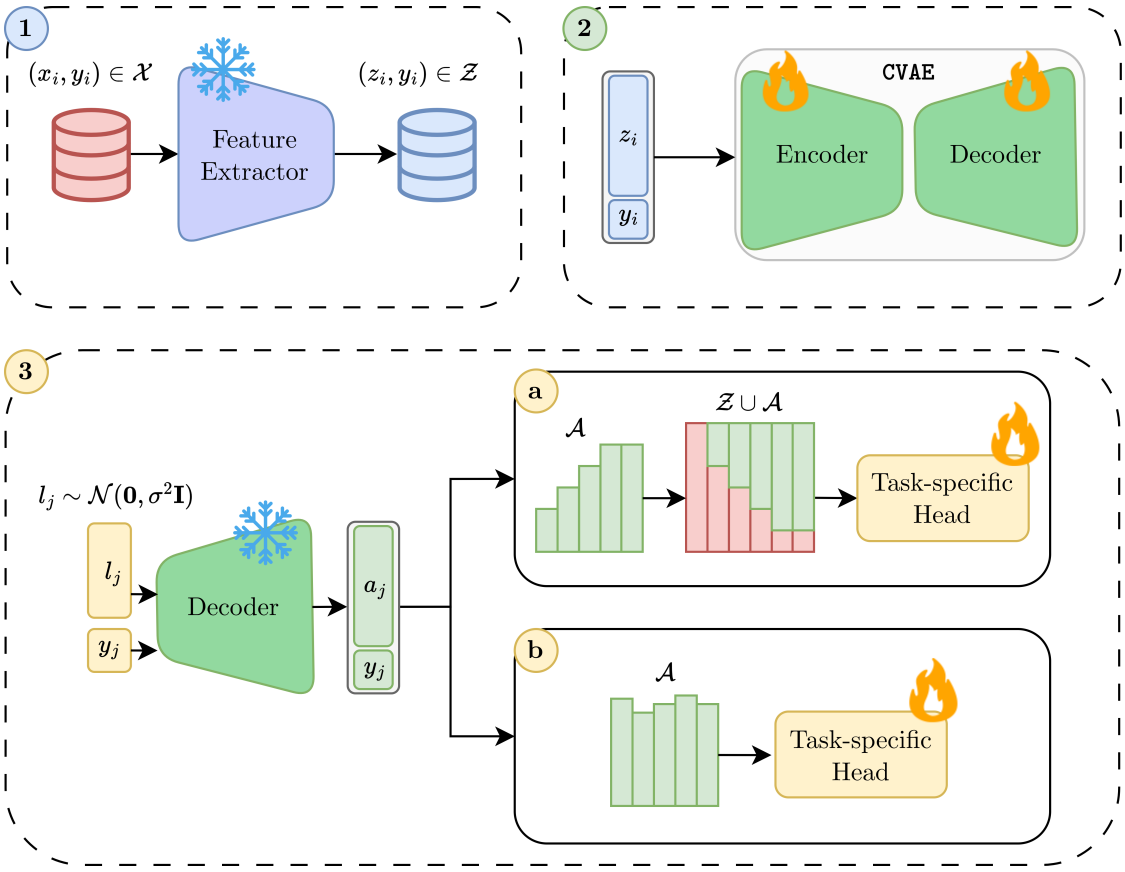

Figure 1: Overview of the proposed methods for long-tailed classification and data anonymization using Conditional Variational Autoencoders (CVAEs). **(1)** Features $\mathcal{Z}$ are extracted from the input data $\mathcal{X}$ using a pre-trained feature extractor. **(2)** The extracted features z_i, along with their corresponding labels y_i, are used to train a CVAE. **(3)** During the generation phase, new samples (a_j, y_j) are generated by sampling $l_j \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$ in the latent space of the CVAE and decoding it conditioned on label y_j. These generated samples form dataset $\mathcal{A}$, that can be combined with the original data $\mathcal{Z}$ for long-tailed classification **(a)** or used independently for data anonymization **(b)**.

2 Background and Related Work

2.1 Foundation Models

2.1.1 Definition

The term 'foundation model' was coined in a report by the Stanford Center for Research on Foundation Models, which defines it as "any model that is trained on broad data (generally using self-supervision at scale) that can be adapted (e.g., fine-tuned) to a wide range of downstream tasks" (Bommasani et al., 2022). Their significance can further be characterized by a trend towards *homogenization*: their consolidation in numerous application domains. For example, Bommasani et al. (2022) argue that training language models in a self-supervised fashion became the norm in the field of Natural Language Processing, rather than just a sub-discipline, after the appearance of foundation models, such as BERT (Devlin et al., 2019), GPT-2 (Radford et al., 2019), RoBERTa (Liu et al., 2019a), or T5 (Raffel et al., 2019) in 2019. They describe the beginnings of similar homogenization trends for other domains including vision, tabular data, computational biology, and reinforcement learning.

Foundation models do not only homogenize the training methods, such as the combination of language models and self-supervised learning at scale, but also the models themselves. For example, GPT-3 represents a standardized approach to building highly capable language models, demonstrating how the architecture and scale can be uniformly applied across various tasks within a domain. This stands in contrast to earlier paradigms in deep learning, which homogenized specific architectures such as Convolutional Neural Networks (CNNs), and in machine learning, which standardized learning algorithms such as logistic regression (Bommasani et al., 2022).

Another defining feature of foundation models is the *emergence* of functionalities that the model has not been explicitly trained to exhibit. An example of emergence is in-context learning, the ability of language models such as GPT-3 (Brown et al., 2020) to be adapted via natural language prompting, enabling them to learn from context by analogy (Bommasani et al., 2022).

Finally, foundation models can be characterized on a technical level. At this level, Bommasani et al. (2022) highlight the importance of both transfer learning and scale. Transfer learning in the context of foundation models refers to pre-training on a surrogate task (e.g., language modeling, where the task is to predict the next word in a sentence) and subsequent adaptation for a specific downstream task (e.g., text classification, sentiment analysis, or chatbots fine-tuned using Reinforcement Learning from Human Feedback (RLHF)). The scale of foundation models, such as GPT-3 with its 175 billion parameters, is made possible by improvements in computer hardware, the development of the Transformer architecture (Vaswani et al., 2017), and an abundance of training data (Bommasani et al., 2022). The resulting capacity of foundation models to learn from large quantities of often unlabeled data and the potential to integrate information from multiple modalities (Radford et al.,

2021; Li et al., 2022; Aghajanyan et al., 2022; Girdhar et al., 2023) leads to a wide range of potential application domains, including healthcare and biomedicine (Zhou et al., 2023; Moor et al., 2023), law, and education (Bommasani et al., 2022). The rest of this subsection describes developments leading to foundation-model-level visual features that we use in this work.

2.1.2 Vision Architectures

Building on the mechanism of self-attention (Bahdanau et al., 2016), the Transformer architecture (Vaswani et al., 2017) enables more flexible and general computation than traditional Multi-Layer Perceptrons (MLPs) and CNNs. This flexibility allows for increased expressivity and better scaling with large amounts of training data (Bommasani et al., 2022). Unlike task-specialized architectures, the Transformer trades off task-specific optimizations for a more general approach, making it suitable for a wide range of applications. Furthermore, the Transformer architecture supports greater parallelization compared to previous sequence model architectures, such as Recurrent Neural Networks (RNNs), leveraging advances in AI accelerators (Shahid and Mushtaq, 2020; Chen et al., 2020), and enabling models of unprecedented scale.

Dosovitskiy et al. (2020) adapted the Transformer architecture for computer vision, resulting in the Vision Transformer (ViT), which was trained on a supervised image classification task. The ViT represents a significant innovation by employing a Transformer-only architecture for computer vision, which can outperform other architectures when trained on large datasets. Figure 2 illustrates the architecture of the ViT.

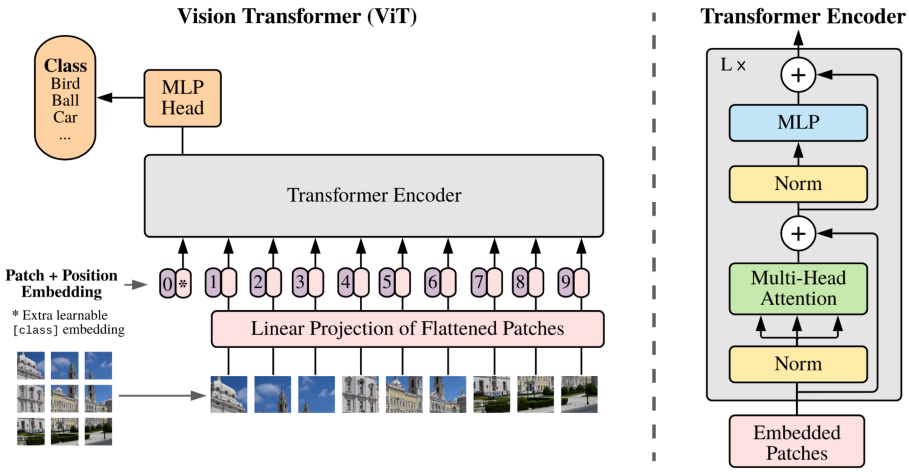

Figure 2: The architecture of the Vision Transformer. The image is divided into patches, which are linearly embedded and combined with position embeddings. These embeddings are processed by a Transformer Encoder comprising multiple layers of multi-head attention and MLP blocks with normalization. The output is then used for classification tasks.This figure is reproduced from Dosovitskiy et al. (2020)

A notable feature of the ViT is its ability to flexibly attend to far-away pixels from the first layer, unlike CNNs where receptive fields are predefined by the architecture. This flexibility is illustrated in Figure 3, which shows the average attention distance across the various layers of the ViT. This illustrates the weaker inductive bias and therefore increased expressivity of the Transformer architecture compared to CNNs.

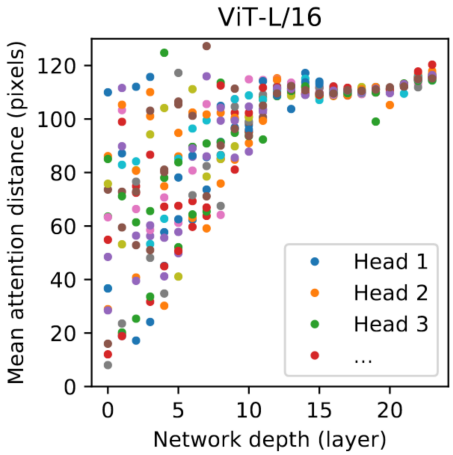

Figure 3: Mean attention distance by attention heads in ViT at different layers. This figure is reproduced from Dosovitskiy et al. (2020)

Contrastive Language-Image Pretraining (CLIP) is one example of a foundation-model architecture that leverages the ViT's computational efficiency with a large pre-training dataset (Radford et al., 2021). As Figure 4 shows, CLIP combines an image encoder, such as a ViT, with a text encoder (also a Transformer) to produce a multimodal embedding space, enabling text-guided visual representation learning. Because CLIP is trained with a contrastive loss that minimizes the cosine distances between pairs of image- and text-embeddings, it can be trained with weakly supervised datasets of image-text pairs. This allows for training the model with larger and potentially cheaper datasets in comparison to high-quality manually-labeled datasets, such as ImageNet (Deng et al., 2009a) or MS-COCO (Lin et al., 2015). As a frame of reference, while ImageNet and MS-COCO contain around 14 million and 328 thousand images, respectively, CLIP has been trained on 400 million (Radford et al., 2021), and OpenCLIP (Cherti et al., 2022) models on up to around 2 billion (Schuhmann et al., 2022) image-text pairs scraped from the internet. Models pretrained with CLIP can learn high-quality, robust, and general representations that are useful across a wide range of computer vision tasks and domains. In addition to its representational capacity, the CLIP architecture supports zero-shot image classification by calculating distances in the embedding space between an image-embedding and text-embeddings of arbitrary concepts. It is also an important building block of the text-conditional image generation system DALL-E (Ramesh et al., 2022). Despite these strengths, CLIP models struggle with more abstract tasks, such as counting, or with generalizing classification performance to image domains that are not represented in the training set (Radford et al., 2021; Cherti et al., 2022).

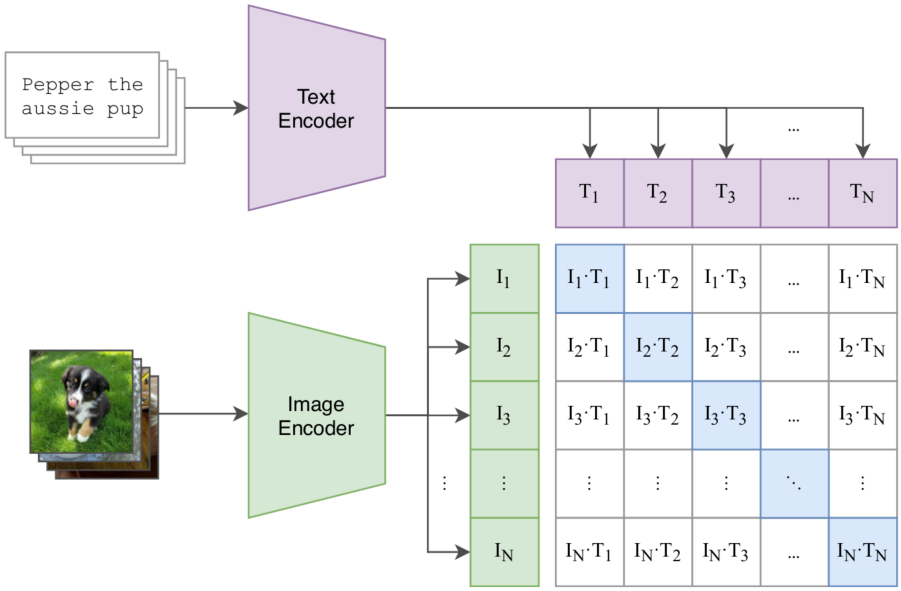

Figure 4: Illustration of the joint embedding space of CLIP, which combines text and image encoders to enable text-guided visual representation learning. This figure is reproduced from Radford et al. (2021)

While CLIP can be efficiently trained on very large datasets, it still relies on reasonably accurate text descriptions to guide the learning of visual features. In contrast, self-supervised training of Transformers has become the norm in language models such as BERT (Devlin et al., 2019). Given the success of self-supervised pre-training in NLP, Caron et al. (2021) extended this idea to Vision Transformers (ViTs). However, the visual domain lacks the inherent sequential structure of language that forms the basis for many common pre-training tasks.

To overcome the challenges posed by the absence of sequential structure in the visual domain, DINO addresses reframes knowledge distillation (Hinton et al., 2015) as a self-supervised pre-training task. As illustrated in Figure 5, various augmentations of the inputs are presented to student and teacher networks. The difference in output features is then computed using a loss function, such as cross-entropy loss. Gradients are propagated back through the student network, and the parameters of the teacher network are updated via an exponentially moving average of the student's network parameters. Oquab et al. (2023) trained DINO models using a larger, curated dataset of 142 million images. The resulting DINOv2 features perform comparably to weakly-supervised methods like CLIP across a variety of tasks and domains, and do not require fine-tuning. Moreover, they learn abstract concepts about parts of objects without explicit training, exhibiting an *emergent* property that is characteristic of foundation models.

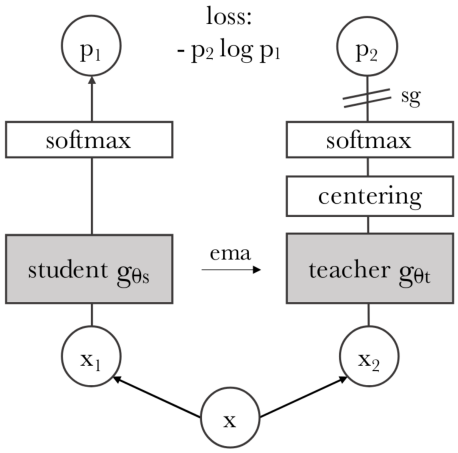

Figure 5: The DINO self-supervised learning framework. This figure is reproduced from Caron et al. (2021)

In summary, self-supervised and weakly supervised pre-training techniques have resulted in powerful foundation models, such as CLIP and DINOv2, which produce highly semantic, robust, and versatile embeddings. These embeddings can be used directly for various downstream tasks without the need for fine-tuning, providing grounds for further exploration on how to best utilize them in diverse applications. In this work, we focus on the applications of long-tailed learning, and privacy, particularly data anonymization within the field of privacy.

2.2 Long-Tailed Learning

Long-tailed learning is a sub-discipline within the field of Machine Learning that is concerned with training datasets that follow a long-tailed class distribution (Zhang et al., 2023). This definition lets us relate long-tailed learning to other common disciplines within Machine Learning. As Figure 6 illustrates, a dataset with a long tail implies a heavy class imbalance, with a few classes containing most of the training samples and most classes consisting of only a few samples. Thus, algorithms dealing with such datasets need to not only deal with bias due to class imbalance but also with problems related to low data availability for most classes. The former problem lets us categorize long-tailed learning as a sub-task of imbalanced learning (He and Garcia, 2009; Wang and Yao, 2012), because a long-tailed distribution is a specific type of imbalanced distribution. The latter problem of limited number of samples for most classes implies that few-shot learning (Wang et al., 2020) must be adequately addressed for an algorithm to successfully learn from a long-tailed dataset. Another perspective from which to view long-tailed learning is out-of-distribution generalization (Liu et al., 2021; Jamal et al., 2020). Here, an algorithm needs to generalize beyond the long-tailed training distribution to balanced, differently imbalanced, or unknown test distributions (Zhang et al., 2022a). Long-tailed learning is relevant for many tasks in computer vision, such as multi-class and/or multi-label classification (Liu et al., 2019b; Wu et al., 2020), object detection (Lin et al., 2017),

and instance segmentation (Zang et al., 2021). In this work, we focus on multi-class, single-label classification, as it provides the most straightforward way to evaluate the approach we present. The following paragraphs provide an overview of the relevant literature in long-tailed learning, offering a basis for the reader to contextualize our proposed approach.

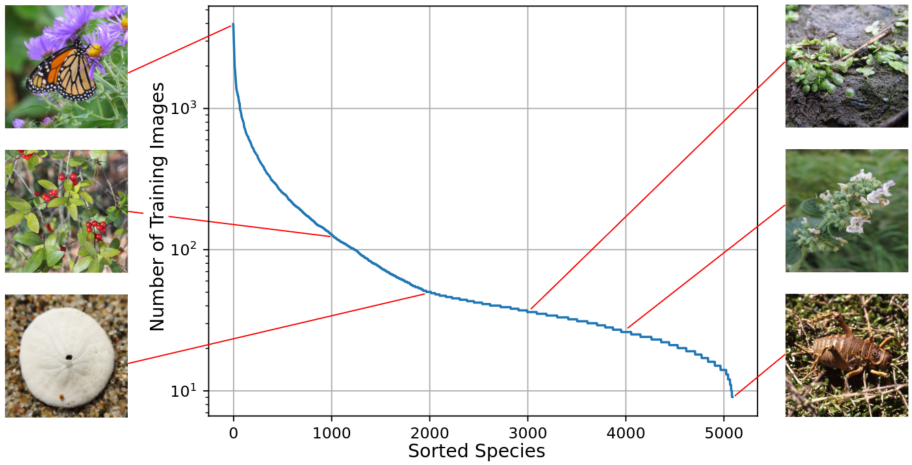

Figure 6: Class distribution of the iNaturalist species classification dataset, illustrating the long-tailed nature of the dataset. The x-axis represents the sorted species, and the y-axis represents the number of training images per species on a logarithmic scale. A few species have a large number of training images, while most species have significantly fewer images. This figure is reproduced from Van Horn et al. (2018).

2.2.1 Cost-Sensitive Learning

The commonly used cross-entropy loss is heavily influenced by class imbalance in training sets, leading to biased classifiers. Several methods have been proposed to solve this issue by adapting and re-balancing the loss values for different classes. For example, focal loss (Lin et al., 2017) lowers or increases the loss for well-classified or less well-classified samples, respectively. LDAM loss (Cao et al., 2019) enforces larger margins for tail classes than for head classes. Class-balanced losses (Cui et al., 2019) reweight samples inversely to the *effective* number of samples in their classes. Balanced softmax loss (Ren et al., 2020) is derived from explicitly modeling test-time distribution shift and weights logits with label frequencies. Equalization loss (Tan et al., 2020) is based on the idea that in conventional cross-entropy loss, rare classes receive discouraging gradients from frequent classes more often than the other way around. It therefore ignores discouraging gradients for rare classes from samples of frequent classes. These techniques are both computationally light and often perform well on benchmark datasets. However, they do not directly address the problem that information about the tail classes is missing in long-tailed datasets. To mitigate this drawback, they can in principle be combined with data augmentation techniques, such as our approach.

2.2.2 Re-Sampling

A very straightforward method to reduce classifier bias is random over- or under-sampling of head classes or tail classes, respectively. When training a deep neural network instead of only a classifier, oversampling can in some cases be beneficial (Buda et al., 2018), but can also lead to overfitting on tail classes and hinder the learning of generalizable features (Zhou et al., 2020; Kang et al., 2019). Instead of re-sampling from tail classes until class balance is achieved, adjusting the sampling frequency via square-root sampling or progressively balanced sampling (Kang et al., 2019), or via monitoring model training (Feng et al., 2021), can mitigate these problems. Also meta-learning has been employed to this end (Ren et al., 2020; Zang et al., 2021), but requires validation sets. In our setting, with no validation set and frozen features, training a classifier by simply oversampling from tail classes until the classes are balanced can achieve competitive results (Kang et al., 2019). Yet, the problem of missing diversity in tail classes remains.

2.2.3 Data Augmentation and Generative Modeling

Data augmentation and generation techniques transform or generate training data to increase the diversity and information in tail classes. Here, it is useful to differentiate between techniques that operate at the level of the input (i.e. images) and the level of the features. At the input level, a popular method is Mixup (Zhang et al., 2018), which can be beneficial for learning representations, but harmful to the performance of classifiers in long-tailed settings (Zhong et al., 2021). While techniques of augmentation (Kim et al., 2020; Chou et al., 2020) and generation (Dablain et al., 2022; Fajardo et al., 2021) at the input level exist, the complexity in the pixel space is significantly higher compared to the feature space. Furthermore, feature-level augmentations are more relevant to our setting of frozen feature encoders.

In the feature space, head classes and tail classes differ in terms of distribution and intra-class diversity (Liu et al., 2020). A common group of augmentations transfers information from head classes to tail classes, often by estimating the variation of head class features (Yin et al., 2019; Liu et al., 2020; Wang et al., 2021), or with class activation maps (Chu et al., 2020). However, they are designed for deep neural network training and involve learning features, not for training a classifier on features of a frozen encoder. FASA (Zang et al., 2021) and MetaSaug (Li et al., 2021) both generate new data points in feature space based on class-wise feature statistics, but rely on the presence of validation sets.

In contrast, the traditional re-sampling-based methods SMOTE (Chawla et al., 2002) and ADASYN (He et al., 2008), as well as the more recent adaptation of Mixup, Remix (Chou et al., 2020), can be directly applied to the embeddings of frozen feature encoders without a validation set. In contrast to our method, they do not explicitly model the class distributions with conditional generative models. Instead, SMOTE (Synthetic Minority Over-sampling Technique) generates synthetic samples for minority classes by interpolating between existing minority samples. This helps

to balance the dataset and improve the performance of classifiers on imbalanced data. ADASYN (Adaptive Synthetic Sampling) is an extension of SMOTE that focuses on generating synthetic samples for minority class instances that are harder to learn. It adaptively shifts the decision boundary toward the difficult samples, improving the classifier's performance on these challenging cases. Remix is a Mixup-based (Zhang et al., 2018) data augmentation method that combines samples from different classes to create new training samples. It sets the mixing factors independently for samples and labels, allowing to set higher weights to minority classes. A method similar to ours is the delta-encoder (Schwartz et al., 2018), which trains the bottleneck of an autoencoder to represent the difference between two examples from the same class. In contrast, we use class labels to condition the decoder, use a Variational Autoencoder, and focus on long-tailed classification, rather than few-shot recognition.

2.3 Privacy

Privacy in machine learning is a paramount concern, especially when dealing with sensitive data such as medical records, financial transactions, or personal identifiers. The increasing capabilities of deep learning models have heightened these concerns as they can inadvertently memorize and reveal sensitive information. This vulnerability has led to the development of various attack vectors, such as membership inference attacks (Shokri et al., 2017), where an adversary can infer whether a given sample was part of the model's training data. Protecting against such privacy breaches is crucial for maintaining user trust and complying with data protection regulations, such as the EU's General Data Protection Regulation, and the Health Insurance Portability and Accountability Act in the United States.

Several strategies have been proposed to address privacy concerns in machine learning. Differential privacy involves adding noise to the data or the learning process to ensure that the presence or absence of any single data point does not significantly affect the outcome, providing a mathematical guarantee of privacy (Dwork, 2006). However, this technique can degrade model performance, and its applicability varies across different model architectures (Abadi et al., 2016; Ziller et al., 2021). Federated learning allows models to be trained across multiple decentralized devices holding local data samples, without exchanging the data itself, significantly reducing the risk of data leakage but introducing complexity in implementation and management (Kairouz et al., 2021). Homomorphic encryption allows computations to be performed on encrypted data, producing encrypted results that can only be decrypted by the data owner (Acar et al., 2018). While it offers strong privacy guarantees, it is computationally intensive and currently impractical for large-scale machine learning tasks (Lee et al., 2022). Generative modeling offers another avenue for privacy-preserving machine learning, with approaches based on Generative Adversarial Networks (GANs) and autoencoders being particularly prominent in recent years.

Generative Adversarial Networks (GANs) consist of two networks, a generator and a discriminator, that are trained together. The generator creates synthetic data, while the discriminator attempts to distinguish between real and synthetic data (Goodfellow

et al., 2014). This adversarial process leads to the generation of high-quality synthetic data that can effectively anonymize the original dataset. GANs have been used to generate synthetic datasets that protect privacy in various applications, including medical imaging and healthcare (Beaulieu-Jones et al., 2019; Choi et al., 2017), face anonymization (Hukkelås et al., 2019; Wu et al., 2019), and privacy-preserving auto-driving (Xiong et al., 2019). Differentially private GAN models have also been proposed (Xie et al., 2018; Jordon et al., 2018).

Autoencoders are encoder-decoder architectures that learn discriminative features by first compressing the input into a lower-dimensional latent space and then reconstructing it. Variants of the autoencoder have been employed in the anonymization of faces (Nousi et al., 2020), speakers (Chouchane et al., 2023; Shamsabadi et al., 2022; Espinoza-Cuadros et al., 2020), text (Weggenmann et al., 2022), and sensor data (Malekzadeh et al., 2019; Hajihassnai et al., 2021; Malekzadeh et al., 2018). Some of these works introduce differentially private versions of autoencoders (Chouchane et al., 2023; Shamsabadi et al., 2022; Weggenmann et al., 2022). Of particular relevance to our work are Conditional Variational Autoencoders (CVAEs) (Sohn et al., 2015), which extend traditional VAEs (Kingma and Welling, 2013) by conditioning on additional information, such as class labels, during the training and generation processes. Hajihassnai et al. (2021) conditioned CVAEs on private attributes of sensor data and modified these attributes during generation. They additionally employed adversarial training to make the latent representation invariant to the private attributes. In contrast, we condition the CVAE on class labels and train it without adversarial objectives.

3 Methods

3.1 Overview and Notation

As argued in section 2.1, the semantic richness of foundation model embeddings makes them suitable features for various downstream tasks without the necessity to fine-tune the feature encoder. Consider a training set $\mathcal{X} = \{x_i, y_i\}, i \in \{1, \ldots, n\}$ of images and their labels, with $n = \sum_{j=1}^{C} n_j$ training samples and C classes. Let f be the frozen feature encoder of a vision foundation model. Then $z_i = f(x_i)$ denotes the embedding of image x_i in the feature space of f, and $\mathcal{Z} = \{z_i, y_i\}, i \in \{1, \ldots, n\}$ is the set of all embeddings and their respective labels of the training set $\mathcal{X}$. Let g be a classifier with parameters θ. Training g on the embedding training set $\mathcal{Z}$ allows the classifier in combination with the feature encoder to predict the label of any given image x_i, such that the prediction $\hat{y}_i = g(f(x_i))$.

As illustrated in Figure 1, our approach comprises three main stages: feature extraction, CVAE training, and data augmentation.

1. **Feature Extraction**: Given a training set $\mathcal{X}$, we use the frozen feature encoder to extract embeddings z_i, resulting in the embedding set $\mathcal{Z}$.
2. **CVAE Training**: We train a CVAE on the embedding set $\mathcal{Z}$. The CVAE consists of an encoder and a decoder. The encoder probabilistically maps the embeddings z_i, conditioned on their labels y_i, to the CVAE's latent space, resulting in latent vectors l_i. The decoder reconstructs the embeddings from the latent space, also conditioned on the labels.
3. **Data Augmentation**: This stage has two primary applications: (a) long-tailed learning and (b) data anonymization.
   (a) **Long-Tailed Learning**: We generate new embeddings $\mathcal{A}$ by sampling from $\mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$ in the latent space of the trained CVAE. These generated embeddings, combined with the original embeddings $\mathcal{Z}$, form an augmented, rebalanced training set. This augmented set is used to train a classifier g. The generated embeddings aim to increase intra-class diversity, particularly for underrepresented classes in long-tailed datasets.
   (b) **Data Anonymization**: Alternatively, we can use the CVAE to generate anonymized embeddings $\mathcal{A}$ by sampling from the latent space, similarly to (a). These embeddings $\mathcal{A}$ can then be used to train a classifier g and are different enough from $\mathcal{Z}$ to preserve the privacy of the original data.

The entire process is depicted in Figure 1, where each stage is visualized, demonstrating the flow from raw images to augmented embeddings and the final classifier training. In summary, the proposed method leverages the semantic richness of foundation model embeddings and CVAE-based generative augmentation to address both long-tailed learning and data anonymization. The remaining parts of this section describe the method in more detail.

3.2 Conditioning Variational Autoencoders on Class Labels

Because the embeddings of foundation models are rich in semantic information and at the same time lower-dimensional than most images, training generative models on them is more efficient than on the images themselves. Therefore, we generate additional embeddings in the feature space. To this end, we train a Conditional Variational Autoencoder (CVAE) on the embedding dataset $\mathcal{Z}$ of a given training dataset $\mathcal{X}$. We first describe the formulation and training of the CVAE, before detailing its use for long-tailed classification and privacy-preserving classification in the following subsections.

Similarly to the conventional Variational Autoencoder (VAE) (Kingma and Welling, 2013), the CVAE takes as input a training sample and probabilistically maps it to a latent space following a prior which is usually a multivariate isotropic normal distribution. It then attempts to reconstruct the sample from the latent representation. The stochastic mapping into the latent space encourages the VAE to learn a continuous latent space without large gaps. Additionally, the predetermined prior distribution is a well-defined distribution that can be used for sampling from the latent space after training. In this work, we furthermore condition the VAE (Sohn et al., 2015) on the class labels of the training samples in order to generate samples of arbitrary class membership.

We adapt the variational lower bound from Sohn et al. (2015) to our scenario, such that

$$\log p_\phi(z \mid y) \geq -D_{\mathrm{KL}}(q_\phi(l \mid y, z) \parallel p_\phi(l \mid y)) + \mathbb{E}_{q_\phi(l|y,z)}[\log p_\phi(z \mid y, l)], \tag{1}$$

where z and y are variables representing the embedding of an image and its class label, respectively, and l is the corresponding latent variable. Furthermore, $q_\phi(l \mid y, z)$ can be interpreted as the encoder, $p_\phi(l \mid y)$ as the prior, and $p_\phi(z \mid y, l)$ as the decoder. The prior of the latent variable l can furthermore be made statistically independent of the labels y, such that $p_\phi(l \mid y) = p_\phi(l)$ (Sohn et al., 2015).

We parameterize the model with deep neural networks with parameters ϕ, that are optimized via mini-batch gradient descent. Because we train the CVAE end-to-end, we use ϕ to denote all parameters of the model, and do not differentiate between the parameters of the encoder and decoder. This leads to the specification $q_\phi(l \mid y, z) = \mathcal{N}(l; \mu(y, z; \phi), \Sigma(y, z; \phi))$, where μ and Σ are functions learned from data. Because we select the prior to be the multivariate standard normal distribution $p_\phi(l) = \mathcal{N}(\mathbf{0}, \mathbf{I})$, the KL-Divergence in Equation 1 can be computed in closed form. To make the sampling process differentiable, we use the reparameterization trick (Kingma and Welling, 2013). We approximate the expectation in Equation 1 with a Mean Squared Error (MSE) loss and add a scaling factor β to weight the reconstruction and distributional alignment. This results in the following final form of the loss we use to train the CVAEs:

$$L_{CVAE}(y, z; \phi) = MSE(z, z^*) - \beta \cdot KL(q_\phi(l \mid y, z) \parallel p_\phi(l)), \tag{2}$$

where z^* denotes the reconstruction of z.

3.3 CVAEs for Long-Tailed Classification

Although most deep neural networks are trained end-to-end for specific tasks, they notoriously struggle with imbalanced and especially long-tailed training data. This usually takes the form of bias towards head classes and subpar performance on tail classes. One response to this challenge has been to separately train the feature encoder and the classifier (Kang et al., 2019, 2020). Both networks are usually trained on $\mathcal{X}$, but with different objectives for representation learning and the down-stream task. We take this idea one step further and use a pre-trained foundation model as a frozen feature encoder. Foundation models have been trained to produce powerful and versatile features to be used in various settings without fine-tuning (see section 2.1). Here, we explore a way to leverage their embedding spaces using a CVAE in a long-tailed setting.

To more formally introduce the problem of long-tailed classification, consider the training set of images $\mathcal{X}$, as defined in section 3.1. In a long-tailed setting, the classes are sorted by cardinality in decreasing order, such that $n_1 \geq n_2 \geq ... \geq n_C$. It follows from the long-tailed distribution of classes that $n_1 \gg n_C$. Although this definition does not exclusively describe *long-tailed* class distributions, it is common in the literature (Zhang et al., 2023; Kang et al., 2019) and highlights the important property of *class imbalance*. Like Cao et al. (2019), we define the imbalance ratio ρ as the ratio of the cardinalities of the most frequent class to the least frequent class, $\rho = n_1/n_C$.

To approach long-tailed classification, we first train a CVAE on the dataset $\mathcal{Z}$, consisting of foundation model embeddings of the long-tailed dataset of images and labels, $\mathcal{X}$. We then generate more embeddings with the CVAE in order to counter the class imbalance. Due to the probabilistic reconstruction and the KL-divergence term in Equation 2, the latent space of the CVAE is expected to have a smooth multivariate standard normal distribution, making sampling from it a straight-forward process. It is lower-dimensional than the foundation model's embeddings and therefore, a successful encoding by the CVAE should focus on the specific manifold in the foundation model's embedding space on which the training data lie, encoding its most defining features. By providing the labels during training, we not only intend to generate embeddings from arbitrary classes. Importantly, we also aim to create a latent space that encodes the typical variation within classes in the training set. Intuitively, the model does not need to learn to separate the classes in its latent space. Instead, it only needs to learn to represent the variation within them, because the class-specific information is given by the label. We hope that thereby, the model can extrapolate from variation of head classes (i.e. classes with high cardinality, such as the class with cardinality n_1) to what more variation in tail classes (i.e. classes with low cardinality, such as the class with cardinality n_C) could look like. We explore the validity of these assumptions in Sections 4.2 and 4.4.

Because the trained CVAE allows us to generate any number of embeddings for arbitrary classes, numerous possibilities of using them together with $\mathcal{Z}$ to train classifier g exist. We look at this problem from the perspective of combining both resampling and data augmentation, because both are common methods used when training neural networks, particularly on long-tailed datasets (see section 2.2).

During classifier training, the usage of a generated sample can be viewed as a type of data augmentation, where the sample to be augmented is simply replaced by a generated sample of the same class. Specifically, in order to rebalance the classes in $\mathcal{Z}$, we sample each member of the training set $z_c \in \mathcal{Z}$ of class c with probability

$$P_{\text{sample}}(z_c) = \frac{1}{C \cdot n_c}. \tag{3}$$

$P_{\text{sample}}(z_c)$ is inversely related to the number of samples in its class and ensures sampling the same number of samples for each class, in expectation.

Additionally, we replace a sample with a generated one with the probability

$$P_{\text{augment}}(z_c) = S\left(\frac{1}{n_c}\right), \tag{4}$$

where

$$S(x) = \max\left(P_{\min}, \min\left(\frac{x - x_{\min}}{x_{\max} - x_{\min}}, P_{\max}\right)\right) \tag{5}$$

scales its argument to values between 0 and 1, and between a chosen minimum $P_{\min}$ and maximum $P_{\max}$. This increases the diversity of samples from tail classes and avoids overfitting, provided the generative model has successfully learned to generalize from sample diversity in head classes to tail classes.

Algorithm 1 describes this process of resampling and augmentation to create a mini-batch with generated samples. When sampling from the CVAE's latent space, distributions other than the normal distribution are in principle possible. However, since we used the standard normal distribution as a prior when training the CVAE, it has learned to encode inputs close to mean zero and variance one. Therefore, we sample only from normal distributions with mean zero. However, we leave the variance σ^2 as a parameter, in order to control the diversity of the generations. The effects of different variances are explored in Sections 4.3, 4.4, and 4.5.1. Note that sampling from the CVAE's latent space could further be adjusted based on sample or class statistics. Moreover, more involved functions to assign sampling and augmentation probabilities could be useful, such as functions that adapt based on training statistics. This, however, is beyond the scope of this work.

Algorithm 1 Mini-Batch Sampling with Generative Rebalancing

Input: Dataset $\mathcal{Z}$, Mini-batch size M, CVAE decoder `CVAE`, Variance σ^2
Output: Mini-batch $\mathcal{B}$
$\mathcal{B} \leftarrow \emptyset$ ▷ initialize mini-batch
for each $(z, y) \in \mathcal{Z}$ **do**
    $P_{\text{sample}}(z, y) \leftarrow \frac{1}{C * n_y}$ ▷ assign sampling probability based on 3
    $P_{\text{augment}}(z, y) \leftarrow S\left(\frac{1}{n_c}\right)$ ▷ assign augmentation probability based on 4, 5
end for
while $|\mathcal{B}| < M$ **do**
    Sample (z, y) from $\mathcal{Z}$ according to P_{sample} ▷ sample according to P_{sample}
    $a \sim \text{Bernoulli}(P_{\text{augment}}(z, y))$ ▷ determine if augmentation occurs
    if $a = 1$ **then**
        $l \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$ ▷ sample from normal distribution
        $z^* \leftarrow \texttt{CVAE}(l, y)$ ▷ generate z^* using CVAE with label y
        $z \leftarrow z^*$ ▷ set z to generated sample z^*
    end if
    $\mathcal{B} \leftarrow \mathcal{B} \cup \{(z, y)\}$ ▷ add sample to mini-batch
end while
return $\mathcal{B}$ ▷ return the mini-batch

Using the CVAE's decoder to generate new samples can be performed either online (i.e. during training), or offline, where samples are generated and stored before training the classifier. On the one hand, offline generation can in principle make better use of parallel computation by generating large batches of samples at a time, but has higher memory requirements, because the generated samples need to be stored until the classifier has been trained. Online generation, on the other hand, generates samples in memory on the fly, to be discarded after use. This reduces total memory requirements and a potential overhead in disk access time, if the generated samples are saved to disk during offline generation. Importantly, online generation is more flexible with respect to the number of samples it generates. This is helpful in cases where the exact number of samples to be generated for each class is unknown before training, such as can be the case during online-, or lifelong learning. For example, if the relative class sizes are unknown before training, they and therefore also the number of samples to be generated can be estimated based on mini-batch statistics. Similarly, if the total number of training epochs is not known before training, online generation guarantees to generate enough samples. We use online generation in our implementation and experiments.

3.4 CVAEs for Data Anonymization

In this subsection, we propose two methods for using CVAEs to anonymize datasets. The first method generates a persistent anonymized replica of the data, while the second method dynamically creates new data without the need for persistent storage.

Our initial method for anonymizing datasets involves producing and storing synthetic feature vectors that maintain the original dataset's size and class distribution. This anonymization method, outlined in Algorithm 2, utilizes a pre-trained CVAE decoder, a specified number N of synthetic samples to be created, and a categorical distribution K that mirrors the class probabilities of the original dataset. This ensures that the synthetic dataset retains class proportions similar to the original dataset. When sampling from the CVAE's latent space, we use a standard normal distribution here, in contrast to the application for long-tailed learning. As the data anonymization method aspires to retain the statistical properties of the original dataset, it is sensible to sample from the same distribution that was used as a prior distribution when training the CVAE.

Algorithm 2 Anonymize Dataset with CVAE

Input: CVAE decoder `CVAE`, class distribution K, number of samples N
Output: Anonymized dataset $\mathcal{A}$
$\mathcal{A} \leftarrow \emptyset$ ▷ initialize anonymized dataset
while $|\mathcal{A}| < N$ **do**
    $y \sim K$ ▷ sample class label from C
    $l \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ ▷ sample from standard normal distribution
    $z \leftarrow \texttt{CVAE}(l, y)$ ▷ generate data point
    $\mathcal{A} \leftarrow \mathcal{A} \cup \{(z, y)\}$ ▷ add generated sample to dataset
end while
return $\mathcal{A}$ ▷ return the anonymized dataset

Our second anonymization strategy eliminates the need for persistent datasets by using the pre-trained CVAE decoder to generate new data dynamically during task-specific model training. This method, detailed in Algorithm 3, avoids storing or transmitting large volumes of sensitive data and allows sharing the CVAE decoder to reproduce the training data distribution without direct data sharing. In contexts like federated learning, this enhances security. While federated learning trains models on private data without sharing it, model weights can still reveal data characteristics. Our iterative approach generates anonymized features directly, adding security and reducing risks associated with traditional data-sharing methods.

Algorithm 3 Online Anonymization with CVAE

Input: CVAE decoder `CVAE`, class distribution K, task-specific model `Model`
Output: Trained model `Model`
while training not converged **do**
    $y \sim K$ ▷ sample class label from C
    $l \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ ▷ sample from standard normal distribution
    $z \leftarrow \texttt{CVAE}(l, y)$ ▷ generate data point
    `Model.train_step`(z, y) ▷ train model on generated data point
end while
return `Model` ▷ return the trained model

4 Experiments

4.1 General Setup and Datasets

The experiments focused on training CVAEs using 768-dimensional ViT-B DINOv2 embeddings on various image datasets. We explored the CIFAR100 and CIFAR10 datasets (Krizhevsky and Hinton, 2009) and their long-tailed versions, highlighting their class distributions, and included datasets from MedMNISTv2 (Yang et al., 2023) to provide a comprehensive evaluation of our approach. We trained all CVAEs for 500 epochs with a learning rate of 0.001, utilizing the Adam optimizer. The latent dimension was set to 100, the batch size was 256, and the KL-Divergence part of the loss was weighted with $\beta = 0.01$, unless otherwise specified. These values led to stable training of CVAEs in preliminary experimentations and serve as grounds for fair comparison across datasets, sampling variances, and different sizes of CVAEs. We concatenated the one-hot encoded labels to the inputs for both the encoder and decoder. To generate samples, we generally sampled latent vectors from a multivariate standard normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$, unless otherwise specified. We used the PyTorch library (Paszke et al., 2019) to implement all neural networks and ran all experiments on an NVIDIA A100 GPU.

We used the CIFAR datasets to evaluate the properties of CVAEs' latent spaces and our approach to long-tailed learning. CIFAR100 consists of 100 classes with 500 training images and 100 test images per class, and CIFAR10 comprises 10 classes with 5000 training images and 1000 testing images per class. For the long-tailed versions, we imbalanced the training datasets with an exponentially decaying sample size across classes following Cao et al. (2019). We generally used an imbalance ratio of $\rho = 100$, resulting in the largest class having 100 times more samples than the smallest class. For CIFAR100 LT, the largest class contains 500 images, and the smallest class has only 5 images. In CIFAR10 LT, the largest class has 5000 images, while the smallest has 50 images. Figure 7 provides a visual representation of these class distributions.

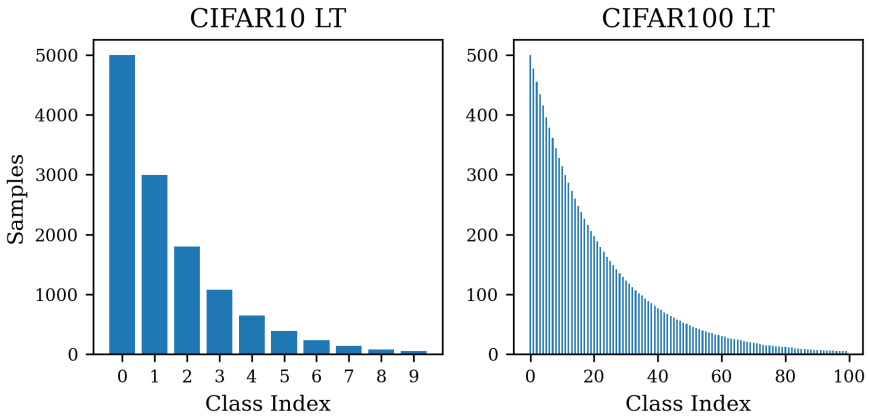

Figure 7: Class distribution in CIFAR10 LT and CIFAR100 LT with imbalance ratio $\rho = 100$.

In addition to the CIFAR datasets, we used the 2D multiclass classification datasets from MedMNISTv2 to evaluate our anonymization approach. These datasets include diverse medical images, each with unique challenges due to varying sizes and complexities. The use of MedMNISTv2 is particularly significant because it tests the CVAEs' ability to generate high-quality, privacy-preserving data in a sensitive domain. Ensuring data privacy in the medical field is crucial, and these datasets allowed us to assess how well the CVAEs can anonymize data without compromising its utility. By incorporating MedMNISTv2, we aim to demonstrate the robustness and versatility of our approach across different types of data and application scenarios. Table 1 gives an overview of the MedMNISTv2 datasets.

Dataset	**Data Modality**	**# Classes**	**# Samples**
PathMNIST	Colon Pathology	9	107,180
DermaMNIST	Dermatoscope	7	10,015
OCTMNIST	Retinal OCT	4	109,309
BloodMNIST	Blood Cell Microscope	8	17,092
TissueMNIST	Kidney Cortex Microscope	8	236,386
OrganAMNIST	Abdominal CT	11	58,830
OrganCMNIST	Abdominal CT	11	23,583
OrganSMNIST	Abdominal CT	11	25,211

Table 1: MedMNISTv2 datasets used for evaluation (Yang et al., 2023).

4.2 Distributions in the Latent Space of VAE and CVAE

In Section 3.3, we argued that, due to the combined effects of the KL-divergence loss and the conditioning on the class labels, the CVAE's latent space should encode intra-class variation independently of the class labels. The CVAE receives all information relevant to class separation as input to both its encoder and its decoder via the labels. Moreover, the KL-Divergence part of its loss function discourages the CVAE to encode samples in ways that deviate from a standard normal distribution. This should result in a latent space that predominantly encodes variation between samples *within* classes. To qualitatively test this assumption, we first trained both a CVAE and a VAE with identical hyperparameters, $\beta = 0.1$ and 2-dimensional latent spaces on DINOv2 embeddings of CIFAR10 LT. The encodings of the CIFAR-10 test set are visualized in Figure 8. While the VAE has learned to separate the classes, there is no such obvious distinction in the CVAE's latent space, where samples of all classes seem to be approximately normally distributed around mean zero and a similar variance for all classes.

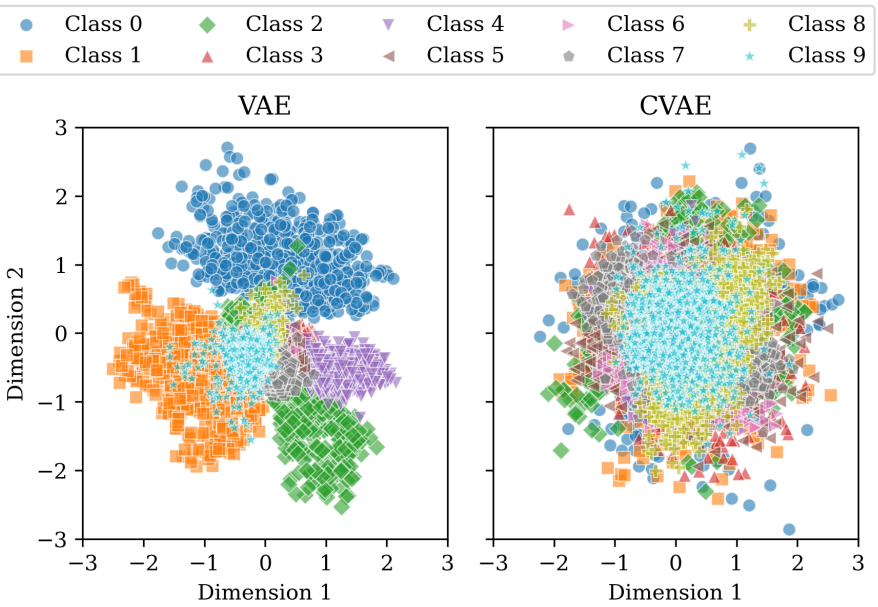

Figure 8: Visualization of CIFAR-10 test set in the latent space of VAE and CVAE.

For quantitative evaluation, we trained a pair consisting of a CVAE and a VAE as described above, and then trained a two-layer dense classifier for each of the generative models on the encoded latents of a random subset comprising 80% of the CIFAR-10 test set. We used the remaining 20% to evaluate the accuracy of the classifier. We repeated the experiment five times and report the average accuracies and reconstruction errors in Table 2. As expected, the classes in the latent space of the VAE are more separable compared to the those in the latent space of the CVAE. However, the CVAE's reconstruction error is generally lower than that of the VAE. This suggests that providing the label during training allows the CVAE to learn

more about the individual and intra-class variation of the training samples, rather than the inter-class variation.

	Accuracy		**MSE**	
	CVAE	**VAE**	**CVAE**	**VAE**
Mean	0.13	0.72	10.70	12.56
Std. Dev.	0.02	0.07	0.10	0.11

Table 2: Separability and reconstruction error in latent space of CVAEs and VAEs. The Accuracy columns show the mean test set accuracy and standard deviation of a classifier trained on latent representations of either a CVAE or a VAE. The Mean Squared error (MSE) columns show the test set reconstruction errors. The results are averaged over 5 runs. Each run consists of training generative models and classifiers.

We further tested with a Kolmogorov-Smirnov test whether the class-wise encoded latents of the CIFAR-10 test set are distributed according to a standard normal distribution for both CVAE and VAE. The results, shown in Table 3, are all highly significant, indicating that neither generative model encodes the test classes according to a standard normal distribution. This can be explained by the relatively low focus on the KL-divergence loss during training ($\beta = 0.1$) and the high power of the tests due to a sample size of 1000 for each class. Importantly, the test statistics for the CVAE are consistently lower than those for the VAE, suggesting that conditioning on the class labels allows the CVAE to more closely approximate a standard normal distribution when encoding its inputs.

	KS Statistic			**KS Statistic**	
label	**CVAE**	**VAE**	**label**	**CVAE**	**VAE**
0	0.14* / 0.14*	0.24* / 0.61*	5	0.17* / 0.13*	0.45* / 0.52*
1	0.15* / 0.15*	0.59* / 0.42*	6	0.2* / 0.16*	0.48* / 0.43*
2	0.088* / 0.15*	0.28* / 0.34*	7	0.13* / 0.15*	0.47* / 0.64*
3	0.14* / 0.16*	0.55* / 0.45*	8	0.3* / 0.14*	0.4* / 0.4*
4	0.17* / 0.14*	0.69* / 0.56*	9	0.21* / 0.19*	0.47* / 0.48*

Table 3: Kolmogorov-Smirnov Statistics statistics for CVAE and VAE normality tests on CIFAR-10 latent space. Each cell shows 'Dimension 1 / Dimension 2' results, with asterisks indicating significance at the 99% confidence level. Lower values indicate a better fit to the normal distribution.

In summary, samples from different classes are less distinguishable in the latent space of CVAEs compared to VAEs. Moreover, CVAEs can more closely satisfy the contraint of a standard normal distribution of data in their latent spaces, while having lower reconstruction error than VAEs. While part of this difference is probably attributable to the additional information that the CVAE is provided with in form of the class labels, as well as the small number of additional parameters due to the concatenation of labels to the input, the results suggest that the latent space of CVAEs encodes mainly intra-class, rather than inter-class variation. In Section 4.4, we try to exploit this property to increase intra-class diversity for tail classes in long-tailed datasets.

4.3 Quality of Generated Embeddings

In this experiment, we compared the quality of embeddings generated by two different CVAE architectures and various sampling variances. The *small* CVAE has a single dense hidden layer of 512 dimensions in both encoder and decoder and around 1 million parameters in total. The *large* CVAE has 6 hidden dense layers (4 with 512 dimensions and 2 with 256 dimensions) in both encoder and decoder, residual connections, and around 3 million parameters. We used DINOv2 embeddings of CIFAR10, CIFAR100, CIFAR10 LT, and CIFAR100 LT as training datasets. For each of the 8 resulting CVAEs, we generated 6 datasets by sampling from a normal distribution in the CVAE's latent space with zero mean and different variances. We evaluated the quality of the generated datasets by calculating the Fréchet Inception Distances (FID) (Heusel et al., 2017) to the test sets of CIFAR10 and CIFAR100. We used the test set of CIFAR10 for datasets generated by models trained on CIFAR10 or CIFAR10 LT and the test set of CIFAR100 for datasets generated by models trained on CIFAR100 or CIFAR100 LT. The FID is commonly used to evaluate the quality of generated images. It compares the distributions of generated data and real data by extracting features from an Inception model and then calculating the Fréchet distance, also known as the Wasserstein-2 distance (Heusel et al., 2017).

The FID is calculated using the following equation:

$$\text{FID} = ||\mu_r - \mu_g||^2 + \text{Tr}(\Sigma_r + \Sigma_g - 2\sqrt{\Sigma_r \Sigma_g}), \tag{6}$$

where μ_r and Σ_r are the mean and covariance of the real data embeddings, and μ_g and Σ_g are the mean and covariance of the generated data embeddings. The square root of the product of the covariance matrices Σ_r and Σ_g is computed using the matrix square root. Here, Tr denotes the trace of a matrix. Since the CVAEs are directly generating feature vectors in the DINOv2 embedding space and not images, we did not compute Inception features. Rather, we compared the generated vectors directly with the DINOv2 features of the appropriate test set. To ensure a fair comparison across different classes, we calculated the FID for each class separately and then took the average. This approach accounts for the class-wise variations in

the data distribution, providing a more robust and comprehensive evaluation of the generative performance of the CVAEs.

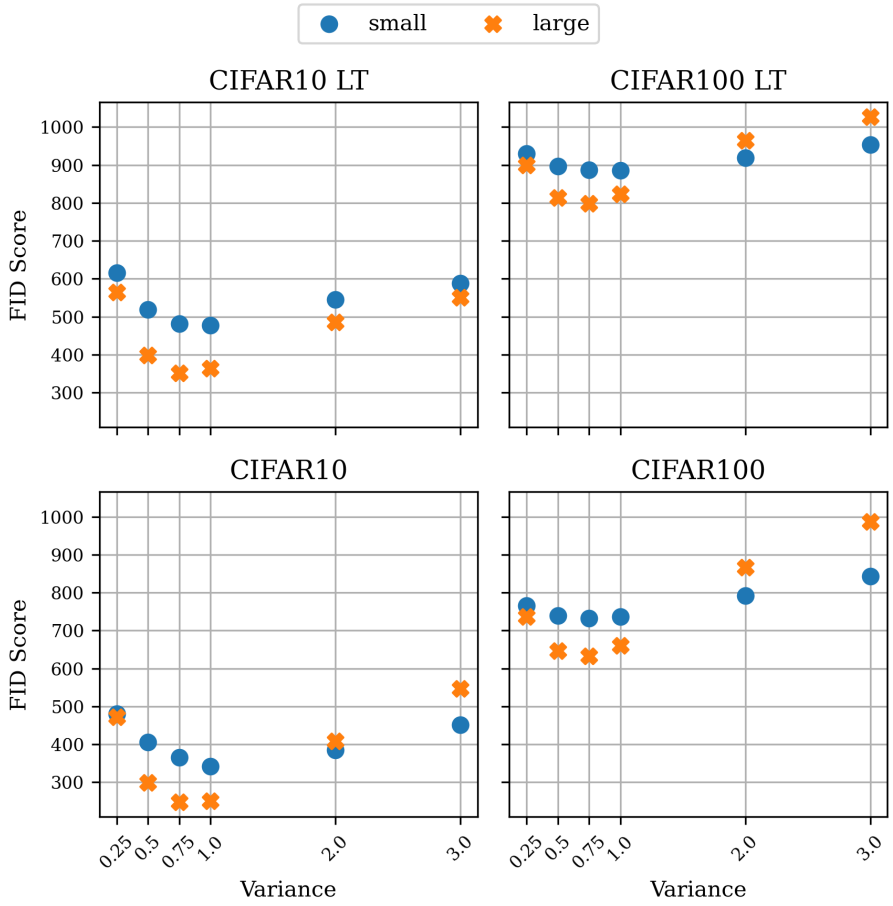

Figure 9: Fréchet Inception Distance (FID) between DINOv2 embeddings of CIFAR test sets and sets of embeddings generated by CVAEs. Plot titles denote the CVAE training set. Sets of generated embeddings are equal to test sets in terms of sample size and distribution across classes. They have been generated by sampling from a normal distribution in the latent space of CVAEs. The x-axis shows the variance used in the sampling process, the y-axis shows the FID score.

The results of our experiments, as illustrated in Figure 9, reveal several key insights into the performance of the two CVAE architectures and the impact of sampling variance. First, the *large* CVAE generally produced lower FID scores compared to the *small* CVAE, indicating a higher quality of generated embeddings. However, this trend tended to reverse at higher variances (above 2), where the *small* CVAE mostly outperformed the *large* CVAE. Furthermore, the FID scores are lowest at moderate variances (0.75 and 1), with both higher and lower variances leading to increased FID scores. This suggests that moderate sampling variance, at values close to what was used for the prior distribution when training the CVAE, strikes a balance between diversity and fidelity in the generated embeddings. Notably, the *large* CVAE appears to be more sensitive to changes in variance, showing greater fluctuation in

FID scores across different variances compared to the *small* CVAE. Additionally, FIDs were generally lower for CVAEs trained on balanced datasets compared to those trained on long-tailed datasets, indicating the advantage of balanced training data. However, variance appears to be the more critical factor in determining FID scores. These findings highlight the importance of carefully tuning the sampling variance and suggest that while the *large* CVAE generally produces higher quality embeddings, it requires more precise control of the variance to maintain this quality.

4.4 Increasing Diversity in Tail Classes

We argued in sections 1 and 3.3 for the need of increased tail class (intra-class) diversity via data augmentation or generation when resampling tail class samples. Here, we explored to what extent a CVAE that has been trained on a long-tailed dataset of foundation model embeddings could be used to meaningfully increase diversity in tail classes. To this end, we first trained a CVAE on the DINOv2 embeddings of the CIFAR100 LT dataset. We used CIFAR100 LT instead of CIFAR10 LT here because it is more representative of long-tailed datasets due to the larger number of classes. We then evaluated intra-class diversity using a coverage metric and the radius of a minimum bounding sphere.

To measure the class-wise coverage of a reference set by samples generated with a CVAE, we used a trained CVAE to generate a dataset with the same number of samples per class as the reference set. For each of the generated samples in each class, we then found its nearest neighbor within the set of samples of the same class in the reference set. Finally, we calculated for each class the proportion of samples of the reference set that is the nearest neighbor to at least one generated sample. We used Euclidean distance to find nearest neighbors. The resulting class-wise coverage metric measures diversity within classes with respect to a specific reference given by the reference dataset. We used the original (balanced) CIFAR100 training set as a reference set in this experiment. Because the CVAE has been trained with a long-tailed subset of this reference set, coverage needs to be interpreted slightly differently for different classes. For the first class, which is identical in the CIFAR100 and CIFAR100 LT datasets, coverage measures how closely the CVAE can reproduce the distribution of samples within a class that it has seen during training. For the other classes, which have increasingly fewer samples in the CIFAR100 LT dataset, coverage measures the proportion of a combination of seen and unseen samples that the generated samples cover.

We measured coverage of the CIFAR100 training set by four datasets for each of the two CVAE architectures. As baseline comparisons, we used the CIFAR100 LT set and the reconstruction of the CIFAR100 training set by the CVAE. The other two datasets were generated by the CVAE by sampling from its latent space with mean 0 and variance 1, or variance 2. Figure 10 presents the results. A first general observation is that coverage for all datasets decreased with increasing class indices, reflecting the decreasing class size of CIFAR100 LT, the CVAE's training set. This is expected even for the generated datasets, because the proportion of unseen samples

increases with higher class indices. Furthermore, the coverage by CVAE-generated datasets tended to be higher than the coverage by CIFAR100 LT, especially for tail classes. Similarly, the datasets generated from randomly sampled latents had higher coverage for tail classes than the reconstructions, suggesting that this method of generation can increase diversity more than simple reconstruction. Overall, the datasets generated this way had flatter curves, suggesting that their coverage of the reference dataset was less dependent on the class size of CIFAR100 LT. A similar trend of flattening can be seen for increasing the variance from 1 to 2. This result was more pronounced in the small CVAE. The large CVAE produced higher coverage by the reconstructed dataset for head classes than the small CVAE, which is less suggestive of generalization ability and more indicative of memorization due to higher capacity.

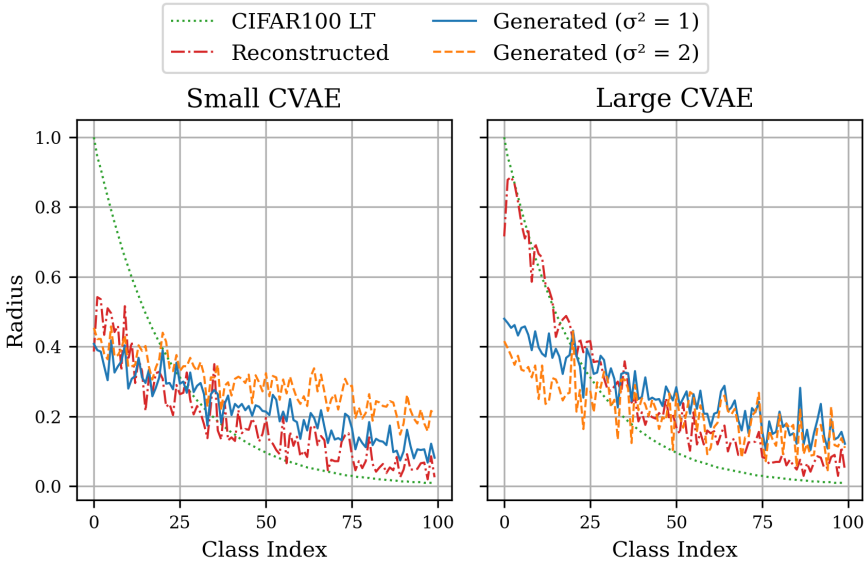

Figure 10: Nearest neighbor coverage of the balanced CIFAR100 training set. CIFAR100 LT: coverage by CIFAR100 LT, the training set of the CVAEs. Generated: coverage by samples generated via random sampling from $\mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$ in the latent space of CVAEs and subsequent decoding. Reconstructed: coverage by reconstructions of the reference set with CVAEs.

While high coverage is an indicator of diversity, it is always with reference to another dataset, and the actual distance to the nearest neighbor might be long, favoring datasets with many and diverse outliers. To assess class diversity from a different angle, we calculated minimum bounding spheres. A minimum bounding sphere is a multi-dimensional sphere with the minimum radius that spans a given set of vectors. We calculated the radius of such a sphere for each class in a dataset. To calculate it, we made an initial estimate of the class center with the feature-wise mean vector and calculated the Euclidean distance to the furthest point from it. We then used SciPy's (Virtanen et al., 2020) optimization package and the L-BFGS-B algorithm

to update the class center to minimize the distance to the furthest point. We used the resulting radius as a metric for intra-class diversity.

Figure 11 shows the radii of minimum bounding spheres for the classes of CIFAR100 and different datasets. The balanced CIFAR100 training set served as a baseline and the radii of its spheres did not change systematically across classes. However, the spheres of the imbalanced CIFAR100 LT dataset decreased with decreasing class size, indicating a reduction in intra-class diversity. We rebalanced the imbalanced dataset with samples generated by the two different CVAE architectures. This generally increased the radii, with higher sampling variance leading to larger radii. The smaller model was more sensitive to changes in the sampling variance than the larger model, and a variance greater than 1 could quickly lead to spheres far larger than those in the balanced dataset. Notably, the reconstructed CIFAR100 dataset had smaller radii compared to the other datasets, indicating that reconstructions had less intra-class diversity. This suggests that while generating new embeddings via random sampling from latent space can increase diversity, reconstructions tend to be more conservative and closely replicate the training data, resulting in smaller bounding spheres.

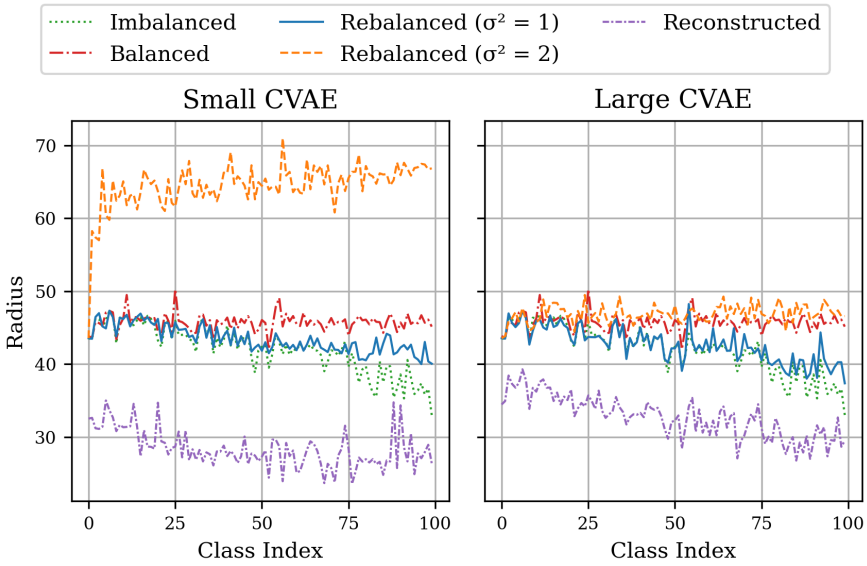

Figure 11: Radii of minimum bounding spheres for classes of CIFAR100. Rebalanced: Rebalanced CIFAR100 LT by samples generated via random sampling from $\mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$ in the latent space of the CVAE and subsequent decoding. Reconstructed: Encoding and subsequent decoding by the CVAE of the balanced CIFAR100 training set.

4.5 Long-Tailed Classification with CVAEs

In this subsection, we examine the impact our generative data augmentation method (Algorithm 1) on the performance of classifiers trained on long-tailed datasets. We conducted experiments using both CIFAR10 LT and CIFAR100 LT datasets, to evaluate the effectiveness of data augmentation using CVAEs in addressing the

challenges of long-tailed distributions. We compared the performance of small and large CVAE architectures across various sampling variances, alongside several augmentation techniques and losses specially designed to tackle long-tailed learning. For the comparisons with other techniques, we used the authors' code where available and easily compatible with our code, or implemented them ourselves. For techniques that required hyperparameters to be set, we used the authors' suggestions or best performing settings found in the original papers for scenarios similar to ours.

The experimental setup for this subsection involved training four CVAEs in total: large and small architectures (as described in 4.3), each trained on CIFAR10 LT and CIFAR100 LT. For embedding generation, we sampled latents from a standard normal distribution, if not specified otherwise. All classifiers consisted of a single dense layer trained with a learning rate of 0.001 using the Adam optimizer for 10 epochs, ensuring convergence and fair comparison. Besides evaluating top-1 accuracy, we report the accuracy of *many-shot* classes (over 100 training samples), *medium-shot* classes (20 to 100 training samples), and *few-shot* classes (under 20 training samples), following Liu et al. (2019b) and Wang et al. (2021).

4.5.1 Effect of Architecture and Sampling Variance

To assess the effect of CVAE architecture and sampling variance on long-tailed classification accuracy, we trained five linear classifiers for each combination of CVAE architecture and sampling variance. We averaged the accuracies over five runs for both CIFAR10 LT and CIFAR100 LT.

For CIFAR100 LT (Table 4), the large CVAE generally outperformed the small CVAE in overall and few-shot accuracies, except at higher variances (2 and 3). The large CVAE exhibited slightly lower accuracy in the many-shot category and showed no clear differences in the medium-shot category. The large CVAE was more sensitive to changes in variance than the small CVAE. The highest overall accuracies were observed at moderate variances of 0.71 and 1. They went hand in hand with the highest or second highest few-shot accuracies and highest medium-shot accuracies, highlighting the importance of high tail class accuracy in long-tailed learning when the test set is balanced. Changes in variance had the largest effect on few-shot classes, followed by medium-shot classes, and the smallest effect on many-shot classes.

	Small				Large			
$\boldsymbol{\sigma^2}$	**O**	**MA**	**ME**	**F**	**O**	**MA**	**ME**	**F**
0.25	83.64	93.09	85.02	71.47	83.67	**93.17**	84.69	71.83
0.5	83.91	93.31	84.85	72.28	84.83	92.87	85.32	75.21
0.75	**84.42**	93.20	**85.22**	73.63	**85.67**	92.37	**85.43**	**78.37**
1	84.22	**93.37**	85.12	72.89	84.62	91.95	84.89	76.03
2	84.02	93.31	84.74	72.74	83.26	92.19	84.02	72.33
3	84.19	93.14	84.42	**73.83**	83.70	92.74	84.60	72.50

Table 4: Mean accuracies for **Small** and **Large** CVAE architectures and sampling variances $\boldsymbol{\sigma^2}$ on CIFAR100 LT with $\rho = 100$, averaged over 5 runs. **O**: Overall, **MA**: Many-shot classes, **ME**: Medium-shot classes, **F**: Few-shot classes.

A similar pattern of results was observed for CIFAR10 LT (Table 5). Note that there are no few-shot classes in CIFAR10 LT at an imbalance rate of $\rho = 100$, because the smallest class has 50 samples and we defined few-shot classes to have fewer than 20 samples. The large CVAE generally outperformed the small CVAE in overall and medium-shot accuracies, while having slightly lower accuracies in the many-shot category. The large CVAE was more sensitive to changes in variance than the small CVAE. Changes in variance had a larger effect on medium-shot classes than on many-shot classes. The highest overall accuracies were observed at variance levels of 0.75 and 1, which also had the highest or second highest medium-shot accuracy.

	Small			Large		
$\boldsymbol{\sigma^2}$	**O**	**MA**	**ME**	**O**	**MA**	**ME**
0.25	97.61	98.37	94.58	97.64	98.36	94.75
0.5	97.72	98.38	**95.05**	97.80	98.37	95.52
0.75	**97.77**	**98.46**	95.02	98.05	**98.42**	96.60
1	97.72	98.39	**95.05**	**98.11**	98.35	**97.19**
2	97.58	98.32	94.62	97.81	98.21	96.18
3	97.68	98.34	95.01	97.73	98.25	95.65

Table 5: Mean accuracies for **Small** and **Large** CVAE architectures and sampling variances $\boldsymbol{\sigma^2}$ on CIFAR10 LT with $\rho = 100$, averaged over 5 runs. **O**: Overall, **MA**: Many-shot classes, **ME**: Medium-shot classes, **F**: Few-shot classes.

4.5.2 Comparison with Other Resampling and Augmentation Methods

We compared our CVAE-based data augmentation method with several baselines: no augmentation, class-balanced sampling (CBS), which adjusts the sampling frequency to ensure minority classes are sampled as often as majority classes, SMOTE (Chawla et al., 2002), ADASYN (He et al., 2008), and Remix (Chou et al., 2020) (see Section 2.2 for descriptions of these methods). These comparisons were conducted on the CIFAR100 LT and CIFAR10 LT datasets (Table 6). For each method and dataset, accuracies were averaged over five runs.

Our method consistently achieved better overall accuracy than Baseline, CBS, SMOTE, and ADASYN for both datasets. It showed similar performance to Remix, with the best overall performance observed when combined with Remix. High accuracies in few-shot and medium-shot groups corresponded to high overall accuracies. Our method had the highest accuracies among single methods for the low-shot groups (few-shot in CIFAR100 LT and medium-shot in CIFAR10 LT).

	CIFAR100 LT				**CIFAR10 LT**		
	O	**MA**	**ME**	**F**	**O**	**MA**	**ME**
Baseline	80.47	**93.80**	82.30	63.40	96.88	98.18	91.67
CBS	83.74	93.02	85.18	71.70	97.78	<u>98.43</u>	95.19
SMOTE	83.31	<u>93.42</u>	84.10	71.04	97.69	98.39	94.87
ADASYN	82.87	93.33	83.78	70.07	97.51	98.35	94.14
Remix	84.91	92.65	<u>85.60</u>	75.42	<u>98.08</u>	**98.51**	96.35
Ours	<u>85.52</u>	91.61	85.54	<u>78.62</u>	98.02	98.29	<u>96.93</u>
Ours + Remix	**85.84**	90.73	**85.61**	**80.57**	**98.15**	98.10	**98.38**

Table 6: Mean accuracies for different resampling and data augmentation techniques on CIFAR100 LT and CIFAR10 LT with $\rho = 100$, averaged over 5 runs. Best and second best results per column are in bold and underlined, respectively. **O**: Overall, **MA**: Many-shot classes, **ME**: Medium-shot classes, **F**: Few-shot classes.

4.5.3 Comparison with Loss Functions for Long-Tailed Learning

We evaluated the performance of our generative data augmentation method in combination with different loss functions designed for long-tailed learning (see Section 2.2): Balanced Softmax (Ren et al., 2020), Class Balanced Softmax (Cui et al., 2019), Focal Loss (Lin et al., 2017), Equalization Loss (Tan et al., 2020), and LDAM Loss (Cao et al., 2019). Accuracies were averaged over five runs for both CIFAR100 LT and CIFAR10 LT.

For CIFAR100 LT (Table 7), our method consistently improved the overall accuracy of the loss functions, except for Balanced Softmax, which had a higher overall accuracy without our method. However, Balanced Softmax accuracy on few-shot samples improved drastically from 79.35 to 85.58 when used with our method. Class Balanced Softmax performed best among the loss functions when combined with our method.

	O	**MA**	**ME**	**F**
Softmax Cross Entropy	80.59	**93.87**	82.30	63.72
+ Ours	**85.40**	91.71	**85.68**	**77.97**
Balanced Softmax	**86.26**	**92.26**	**86.38**	79.35
+ Ours	84.09	82.83	84.02	**85.58**
Class Balanced Softmax	83.17	**93.34**	83.78	71.03
+ Ours	**85.97**	89.43	**84.94**	**83.19**
Focal	78.86	**93.10**	81.26	60.15
+ Ours	**84.12**	90.67	**84.11**	**76.75**
Equalization	82.41	**93.83**	82.04	69.92
+ Ours	**84.52**	88.76	**82.19**	**82.30**
LDAM	77.14	**90.44**	79.54	59.50
+ Ours	**80.65**	89.26	**80.66**	**70.90**

Table 7: Mean accuracies for different loss functions with CVAE data generation (+Ours) and without CVAE data generation on CIFAR100 LT with $\rho = 100$, averaged over 5 runs. Best results between the two methods are in bold. **O**: Overall, **MA**: Many-shot classes, **ME**: Medium-shot classes, **F**: Few-shot classes.

For CIFAR10 LT (Table 8), a similar pattern was observed. Additionally, in contrast to CIFAR100 LT results, adding our method also improved many-shot accuracy for all losses except Balanced Softmax and LDAM.

	O	**MA**	**ME**
Softmax Cross Entropy	96.89	98.19	91.70
+ Ours	**98.11**	**98.38**	**97.00**
Balanced Softmax	**98.26**	**98.42**	97.63
+ Ours	95.74	94.80	**99.47**
Class Balanced Softmax	97.10	98.14	92.93
+ Ours	**98.13**	**98.21**	**97.78**
Focal	96.46	97.89	90.77
+ Ours	**97.72**	**98.06**	**96.34**
Equalization	96.90	98.15	91.92
+ Ours	**98.09**	**98.34**	**97.06**
LDAM	96.28	**97.52**	91.33
+ Ours	**97.19**	97.45	**96.14**

Table 8: Mean accuracies for different loss functions with CVAE data generation (+Ours) and without CVAE data generation on CIFAR10 LT with $\rho = 100$, averaged over 5 runs. Best results between the two methods are in bold. **O**: Overall, **MA**: Many-shot classes, **ME**: Medium-shot classes.

In summary, our experiments demonstrate that CVAE-generated embeddings effectively augment long-tailed datasets, improving classification performance across CVAE architectures and various sampling variances. Combining CVAE-generated data with other feature augmentation techniques and specialized loss functions further enhances performance, particularly for underrepresented classes in long-tailed distributions.

4.6 Data Anonymization with CVAEs

In this subsection, we explore the effectiveness of CVAEs trained on foundation model embeddings for anonymizing datasets while maintaining data utility. We utilized the 2D multiclass classification datasets from MedMNISTv2 (Yang et al., 2023). These datasets are ideal for our experiments due to their real-world nature and the critical importance of data privacy in the medical domain.

We trained a *large* CVAE for each MedMNISTv2 dataset using the hyperparameters as described in Section 4.1. We then employed Algorithm 2 (see Section 3.4), sampling latents from a standard normal distribution, to generate privacy-preserving datasets

that maintained the original number of samples per class. The evaluation of these generated datasets was twofold: assessing performance and ensuring anonymity.

4.6.1 Performance Evaluation

To evaluate the performance of the generated datasets, we trained a linear classifier with a single dense layer with a learning rate of 0.001 using the Adam optimizer and early stopping based on validation set loss for each dataset. We repeated the training process five times and reported the average metrics: accuracy, F1 score, and Cohen's Kappa. These metrics were chosen for their ability to provide a comprehensive assessment of classifier performance. Accuracy indicates the overall correctness of the classifier, the F1 score balances precision and recall, and Cohen's Kappa measures the agreement between predicted and true labels, adjusting for chance. The F1 score is particularly beneficial for multi-class classification as it provides a single metric that accounts for both false positives and false negatives, thus offering a balanced view of performance across classes. Cohen's Kappa is advantageous in multi-class settings because it adjusts for chance agreement and provides a more balanced evaluation across all classes, offering a more robust measure of the classifier's true performance.

The results, presented in Table 9, demonstrate that the classifiers trained on the generated datasets achieved reasonable performance, often comparable to those trained on the original datasets. For example, the generated dataset for PathMNIST achieved an accuracy of 92.72%, an F1 score of 0.928, and a Cohen's Kappa of 0.916, compared to 93.61%, 0.938, and 0.927 for the original dataset, respectively. However, performance varied across datasets, with some generated datasets exhibiting a more significant drop in performance, such as TissueMNIST, where the generated dataset's accuracy was 55.02% compared to 63.47% for the original.

	Accuracy	F1 Score	Cohen's κ
PathMNIST	93.61	0.938	0.927
Generated	92.72	0.928	0.916
DermaMNIST	83.92	0.834	0.681
Generated	75.77	0.731	0.473
OCTMNIST	83.92	0.836	0.786
Generated	79.14	0.787	0.722
BloodMNIST	98.19	0.982	0.979
Generated	96.01	0.960	0.953
TissueMNIST	63.47	0.621	0.533
Generated	55.02	0.509	0.412
OrganAMNIST	91.81	0.917	0.908
Generated	89.19	0.890	0.879
OrganCMNIST	86.89	0.869	0.852
Generated	81.99	0.818	0.796
OrganSMNIST	76.25	0.758	0.728
Generated	70.29	0.700	0.660

Table 9: Mean test set performance of classifiers trained on embeddings of the original versions and anonymized (Generated) versions of the 2D multi-class MedMNISTv2 datasets, averaged over 5 runs.

4.6.2 Anonymity Evaluation

To assess the anonymity of the generated datasets, we calculated the Euclidean distance from every generated point to the nearest point in the original dataset. The distribution of these distances provides insight into the separation between generated and original data points, thereby indicating the level of anonymity.

Figure 12 shows the distributions of these distances. The plots on the left illustrate distances from points in the generated dataset to the nearest neighbor in the original dataset, while the plots on the right depict distances to nearest neighbors within the original dataset as a reference. Notably, distances from generated embeddings to original embeddings are never zero, indicating that the generated embeddings do not overlap directly with any original embeddings, thus preserving anonymity. Additionally, the distances from generated embeddings to original embeddings are

generally larger than the distances within the original dataset, serving as a reference point for comparison. This indicates that the generated embeddings are sufficiently distinct from the original embeddings, thereby preserving anonymity.

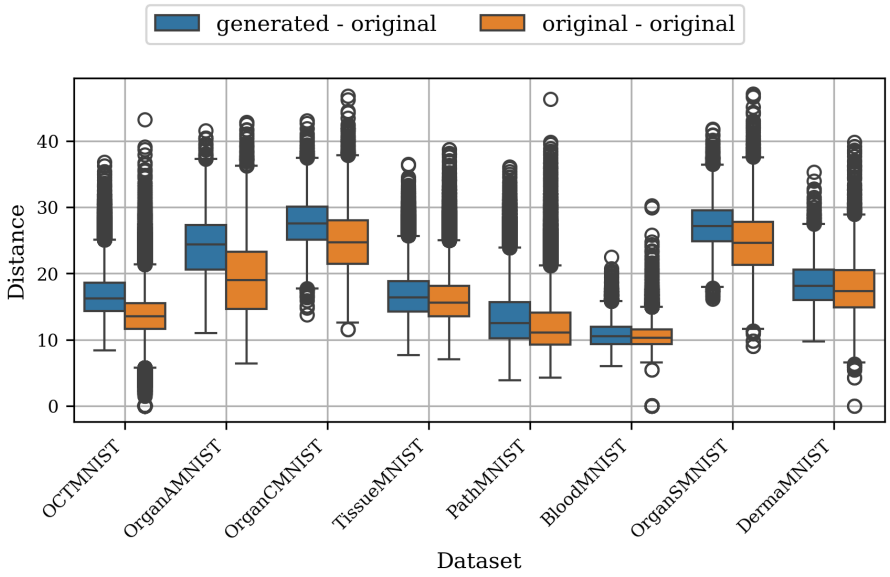

Figure 12: Distributions of distances to nearest neighbors for 2D multiclass datasets from MedMNISTv2. Plots on the left show distances from embeddings in the generated dataset to their nearest neighbors in the original dataset. Plots on the right show distances to nearest neighbors within the original dataset as a frame of reference.

In conclusion, our experiments demonstrate that CVAEs can effectively anonymize datasets, producing synthetic data that maintains reasonable classifier performance while ensuring a significant level of privacy. The variability in performance across different datasets highlights the importance of dataset-specific considerations when applying such anonymization techniques. The distances to the nearest neighbors confirm that the generated data points do not reveal information about individual data points in the original dataset, thus providing a robust privacy-preserving solution for sensitive data.

5 Discussion

5.1 Main Findings and Interpretation

In this thesis, we explored the application of generative data augmentation techniques in the embedding space of vision foundation models to address the challenges posed by long-tailed learning and privacy constraints. The primary objective was to enhance the performance of classifiers in scenarios with imbalanced data distributions and to provide a method for data anonymization without compromising the utility of the data.

Our approach utilized a Conditional Variational Autoencoder (CVAE) trained on the embeddings produced by a vision foundation model. Specifically, we employed the DINOv2 model, a state-of-the-art self-supervised vision transformer, to extract rich and versatile features from images. The CVAE was then conditioned on class labels and trained to generate new embeddings that could augment the existing dataset, particularly focusing on underrepresented classes. The experiments were conducted on several datasets, including CIFAR10, CIFAR100, and MedMNISTv2, with both balanced and long-tailed versions of the CIFAR datasets.

The results from our CVAE-based approach indicated several key findings:

- **CVAE Generated Embeddings:** Our method increases tail class diversity and achieves better quality with larger CVAEs, particularly when sampling from the latent space with a variance of around 1.
- **Long-Tailed Classification:** Our method improves classification performance on long-tailed datasets, evidenced by increased accuracy overall and in tail classes, leading to more balanced performance across the entire class distribution.
- **Privacy Preservation:** The method provides a robust means of anonymizing data, with the generated samples being sufficiently distinct from the original ones to prevent re-identification while retaining the utility for training downstream models.

In terms of the quality of CVAE-generated embeddings, we found that smaller CVAEs are more sensitive to changes in variance for the diversity metrics used, whereas larger CVAEs are more sensitive to variance changes regarding FID quality and accuracy. A possible explanation for these results is that the diversity metrics we used may be insufficient to accurately capture diversity in high-dimensional feature spaces due to the curse of dimensionality and their sensitivity to outliers. Importantly, with the simple and reasonable sampling strategy of using the same distribution for sampling that was used as the prior distribution during CVAE training (standard normal), larger models consistently achieved better results than smaller models and most other variance settings.

For long-tailed classification, our method compares favorably to traditional resampling and feature augmentation methods and shows great promise when combined with Remix augmentation (Chou et al., 2020) and specialized loss functions for long-tailed learning. Specifically, the Balanced Softmax loss function (Ren et al., 2020) had better overall accuracy without our method but improved few-shot accuracy when used in conjunction with our generative data augmentation. This indicates that simply generating samples to rebalance a training set and using a special loss function as if the training set were still long-tailed might not be optimal. Potential solutions to achieve a better trade-off in performance between head and tail classes include weighting generated samples differently from original samples in the loss function, generating different numbers of samples (e.g., not completely rebalancing), and adjusting the class weights used in the Balanced Softmax loss function.

In the context of privacy preservation, our method showed promising results. The distances between generated and original samples were generally larger than those between original samples, indicating effective anonymization. The generated embeddings preserved the statistical properties of the original dataset, ensuring that the utility of the data was maintained while protecting individual data points from re-identification. This approach is particularly advantageous in scenarios where data privacy is paramount, such as in medical applications. The effectiveness of our method is further corroborated by results from a paper by us, currently under review (Di Salvo et al., 2024), which includes evaluations on more datasets and comparisons to the k-Same approach (Newton et al., 2005).

5.2 Contributions to the Field

Our study makes several novel contributions to the fields of generative data augmentation, long-tailed learning, and privacy preservation in machine learning. These contributions enhance the state of knowledge by providing innovative methods and insights that address persistent challenges in these areas.

One of the primary novel contributions is the application of Conditional Variational Autoencoders (CVAEs) in the embedding space of vision foundation models for data augmentation. Previous approaches often focused on augmenting data at the input level (e.g., image pixels) or required expensive learning of feature spaces. Our method leverages the rich, semantically meaningful embeddings generated by foundation models, such as DINOv2 (Oquab et al., 2023), and applies generative modeling techniques directly in this lower-dimensional, feature-rich space. This not only reduces the computational complexity compared to pixel-level augmentation but also capitalizes on the generalization capabilities of pre-trained foundation models, which have been trained on vast and diverse datasets.

A key novelty lies in our approach to addressing long-tailed learning. By generating embeddings for underrepresented classes, our method enhances the diversity within these classes, leading to more balanced datasets and improved classifier performance. The use of CVAEs allows for the generation of samples that capture intra-class

variation, which is crucial for robust learning in long-tailed datasets. By directly increasing the meaningful diversity of the training data, this method goes beyond traditional resampling and reweighting techniques.

Additionally, our study highlights the sensitivity of CVAEs to sampling variance and model size, providing new insights into the optimal conditions for generating high-quality synthetic embeddings. While larger CVAEs generally produced better results, they required conservative sampling strategies from their latent spaces to create samples that contribute positively to the classifier's performance. This finding underscores the importance of balancing model complexity and sampling strategy to achieve the best outcomes in generative data augmentation.

In the domain of privacy preservation, our work introduces a novel method for data anonymization using generative modeling in the embedding space. Unlike traditional anonymization techniques, which often rely on obfuscating or perturbing the data at the input level, our approach generates entirely new embeddings that retain the statistical properties of the original dataset while ensuring privacy. This method effectively prevents re-identification of individual data points and offers a robust solution for scenarios where data privacy is crucial.

5.3 Potential Avenues for Future Research

Our study opens up several promising avenues for future research, each aimed at further enhancing the capabilities and applications of generative data augmentation in the embedding space of vision foundation models. These directions can be explored to address remaining challenges and extend the applicability of our methods.

One potential direction for future research is the exploration of different types of generative models beyond Conditional Variational Autoencoders (CVAEs). While CVAEs have proven effective in our experiments, other generative models such as Generative Adversarial Networks (GANs) or Diffusion Models could offer different advantages. GANs, for instance, are known for generating high-quality samples and could be adapted to work in the embedding space of vision foundation models. Diffusion models are already being successfully trained in the latent space of pretrained autoencoders for the image generation method Stable Diffusion (Rombach et al., 2022). Comparing these models with CVAEs in terms of performance, computational efficiency, and quality of generated embeddings would provide valuable insights.

Another area worth exploring is the use of more specialized foundation model encoders. While we used the DINOv2 model due to its strong performance and generalization capabilities, there are other models, particularly those trained on specific domains such as medical imaging, that could offer enhanced performance for certain applications (Zhang et al., 2022b; Zhou et al., 2023). Investigating how these more specialized encoders interact with our generative data augmentation method could lead to further improvements in performance, especially for domain-specific tasks.

Additionally, future research could focus on refining the sampling strategies used during the generation of embeddings. Our results showed that the variance in the sampling process significantly affects the quality of the generated embeddings. Developing more sophisticated sampling strategies that adaptively adjust the variance based on the characteristics of the data could lead to better results. For example, methods that dynamically adjust the sampling variance during training of a downstream task-specific network might produce embeddings that better balance diversity and fidelity.

The integration of our generative data augmentation method with various loss functions and other augmentation methods designed for long-tailed learning is another promising research direction. While we demonstrated improvements using our method, combining it with advanced loss functions, such as Balanced Softmax loss (Ren et al., 2020), and data augmentation methods, such as Remix (Chou et al., 2020), could yield even better results. Investigating how these methods interact with the generated embeddings and which specific adjustments are needed for optimal performance would be a valuable contribution.

Our approach to privacy preservation through generative modeling also suggests several future research opportunities. One potential direction is to enhance the anonymity of the generated data by developing more sophisticated techniques for ensuring that generated samples are sufficiently distinct from original samples. For instance, implementing mechanisms that enforce a minimum distance between generated and original samples in the embedding space would be straightforward and could further improve privacy guarantees. Additionally, exploring the integration of differential privacy techniques or federated learning approaches with our generative model could provide robust frameworks that combine the strengths of these approaches.

Finally, real-world testing and validation of our methods on larger and more complex datasets, such as iNaturalist (Van Horn et al., 2018) for long-tailed learning, would provide a deeper understanding of their practical applicability. Such experiments could help identify any limitations or challenges that arise in real-world scenarios and lead to the development of more refined and scalable solutions.

5.4 Strengths and Limitations

This study has several strengths that contribute to its impact and validity, as well as some limitations that should be acknowledged.

Next to its practical effectiveness in long-tailed classification and data anonymization, the primary strength of our approach is its generality and flexibility. By leveraging vision foundation models and focusing on the embedding space rather than the raw input space, we developed a generative data augmentation method that is broadly applicable across various domains and datasets. This flexibility makes our approach particularly valuable for practitioners who may not have the resources to fine-tune large models but still need to improve performance on specific tasks.

However, our study also has limitations that should be considered. One limitation is the dependency on the quality of the embeddings produced by the foundation model. While we used DINOv2, which provides robust and versatile features, the performance of our generative data augmentation method is inherently tied to the quality of these embeddings. If the foundation model's embeddings are not well-suited for a particular task or dataset, the effectiveness of our approach may be compromised.

Another limitation is the sensitivity of the CVAE to the variance used in the sampling process. Our results indicated that the performance of the generative model could vary significantly with different sampling variances, especially for larger CVAEs. This sensitivity requires careful tuning and may pose a challenge in practical applications where optimal variance settings are not known a priori. Further research into adaptive sampling strategies could help address this limitation.

Additionally, while our method showed promising results in the datasets we tested, its generalizability to more complex and larger-scale real-world datasets remains to be fully validated. Testing our approach on such datasets would provide a more comprehensive understanding of its scalability and robustness in diverse scenarios.

Finally, our privacy preservation method, while effective, does not offer formal mathematical guarantees like differential privacy (Dwork, 2006). While the generated embeddings are distinct enough to prevent re-identification in our experiments, formal privacy guarantees would provide additional confidence in the robustness of our approach. Integrating differential privacy techniques with our generative model could be a future direction to strengthen the privacy aspects of our method.

6 Conclusion

In this thesis, we proposed a novel approach to address the challenges of long-tailed learning and privacy constraints by leveraging generative data augmentation in the embedding space of vision foundation models. By training Conditional Variational Autoencoders (CVAEs) on the embeddings of pre-trained vision models, we demonstrated an effective method for generating synthetic samples that enhance the diversity of underrepresented classes. Our experiments showed that this technique not only improves classification performance on long-tailed datasets but also provides a robust framework for anonymizing sensitive data, thereby preserving privacy without compromising data utility. The results validate the potential of combining generative modeling with foundation model embeddings to tackle key issues in modern machine learning, offering a scalable and versatile solution for diverse real-world applications. Future work could explore further optimization of CVAE architectures and the integration of additional privacy-preserving mechanisms to enhance the robustness and applicability of this approach.

Bibliography

Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep Learning with Differential Privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, CCS '16, pages 308–318, New York, NY, USA, October 2016. Association for Computing Machinery. ISBN 978-1-4503-4139-4. doi: 10.1145/2976749.2978318.

Abbas Acar, Hidayet Aksu, A. Selcuk Uluagac, and Mauro Conti. A Survey on Homomorphic Encryption Schemes: Theory and Implementation. *ACM Computing Surveys*, 51(4):79:1–79:35, July 2018. ISSN 0360-0300. doi: 10.1145/3214303.

Armen Aghajanyan, Bernie Huang, Candace Ross, Vladimir Karpukhin, Hu Xu, Naman Goyal, Dmytro Okhonko, Mandar Joshi, Gargi Ghosh, Mike Lewis, and Luke Zettlemoyer. CM3: A Causal Masked Multimodal Model of the Internet, January 2022.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural Machine Translation by Jointly Learning to Align and Translate, May 2016.

Brett K. Beaulieu-Jones, Zhiwei Steven Wu, Chris Williams, Ran Lee, Sanjeev P. Bhavnani, James Brian Byrd, and Casey S. Greene. Privacy-Preserving Generative Deep Neural Networks Support Clinical Data Sharing. *Circulation: Cardiovascular Quality and Outcomes*, 12(7):e005122, July 2019. ISSN 1941-7713, 1941-7705. doi: 10.1161/CIRCOUTCOMES.118.005122.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga,

Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the Opportunities and Risks of Foundation Models, July 2022.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.

Mateusz Buda, Atsuto Maki, and Maciej A. Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106: 249–259, October 2018. ISSN 0893-6080. doi: 10.1016/j.neunet.2018.07.011.

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning Imbalanced Datasets with Label-Distribution-Aware Margin Loss. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging Properties in Self-Supervised Vision Transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021.

N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16: 321–357, June 2002. ISSN 1076-9757. doi: 10.1613/jair.953.

Yiran Chen, Yuan Xie, Linghao Song, Fan Chen, and Tianqi Tang. A Survey of Accelerator Architectures for Deep Neural Networks. *Engineering*, 6(3):264–274, March 2020. ISSN 2095-8099. doi: 10.1016/j.eng.2020.01.007.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning, December 2022.

Edward Choi, Siddharth Biswal, Bradley Malin, Jon Duke, Walter F. Stewart, and Jimeng Sun. Generating Multi-label Discrete Patient Records using Generative Adversarial Networks. In *Proceedings of the 2nd Machine Learning for Healthcare Conference*, pages 286–305. PMLR, November 2017.

Hsin-Ping Chou, Shih-Chieh Chang, Jia-Yu Pan, Wei Wei, and Da-Cheng Juan. Remix: Rebalanced Mixup. In Adrien Bartoli and Andrea Fusiello, editors, *Computer Vision – ECCV 2020 Workshops*, pages 95–110, Cham, 2020. Springer International Publishing. ISBN 978-3-030-65414-6. doi: 10.1007/978-3-030-65414-6_9.

Oubaïda Chouchane, Michele Panariello, Oualid Zari, Ismet Kerenciler, Imen Chihaoui, Massimiliano Todisco, and Melek Önen. Differentially Private Adversarial Auto-Encoder to Protect Gender in Voice Biometrics. In *Proceedings of the 2023 ACM Workshop on Information Hiding and Multimedia Security*, IH&MMSec '23, pages 127–132, New York, NY, USA, June 2023. Association for Computing Machinery. ISBN 9798400700545. doi: 10.1145/3577163.3595102.

Peng Chu, Xiao Bian, Shaopeng Liu, and Haibin Ling. Feature Space Augmentation for Long-Tailed Data, August 2020.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-Balanced Loss Based on Effective Number of Samples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9268–9277, 2019.

Damien Dablain, Bartosz Krawczyk, and Nitesh V. Chawla. DeepSMOTE: Fusing Deep Learning and SMOTE for Imbalanced Data. *IEEE Transactions on Neural Networks and Learning Systems*, 34(9):6390–6404, January 2022. ISSN 2162-2388. doi: 10.1109/TNNLS.2021.3136503.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, June 2009a. doi: 10.1109/CVPR.2009.5206848.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. Ieee, 2009b.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, May 2019.

Francesco Di Salvo, David Tafler, Sebastian Doerrich, and Christian Ledig. Privacy-preserving datasets by capturing feature distributions with Conditional VAEs. 2024.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*, October 2020.

Cynthia Dwork. Differential Privacy. In Michele Bugliesi, Bart Preneel, Vladimiro Sassone, and Ingo Wegener, editors, *Automata, Languages and Programming*, pages 1–12, Berlin, Heidelberg, 2006. Springer. ISBN 978-3-540-35908-1. doi: 10.1007/11787006_1.

Fernando M. Espinoza-Cuadros, Juan M. Perero-Codosero, Javier Antón-Martín, and Luis A. Hernández-Gómez. Speaker De-identification System using Autoencoders and Adversarial Training, November 2020.

Val Andrei Fajardo, David Findlay, Charu Jaiswal, Xinshang Yin, Roshanak Houmanfar, Honglei Xie, Jiaxi Liang, Xichen She, and D. B. Emerson. On oversampling imbalanced data with deep conditional generative models. *Expert Systems with Applications*, 169:114463, May 2021. ISSN 0957-4174. doi: 10.1016/j.eswa.2020.114463.

Chengjian Feng, Yujie Zhong, and Weilin Huang. Exploring Classification Equilibrium in Long-Tailed Object Detection. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3397–3406, Montreal, QC, Canada, October 2021. IEEE. ISBN 978-1-66542-812-5. doi: 10.1109/ICCV48922.2021.00340.

Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, CCS '15, pages 1322–1333, New York, NY, USA, October 2015. Association for Computing Machinery. ISBN 978-1-4503-3832-5. doi: 10.1145/2810103.2813677.

Matthew Fredrikson, Eric Lantz, Somesh Jha, Simon Lin, David Page, and Thomas Ristenpart. Privacy in Pharmacogenetics: An {End-to-End} Case Study of Personalized Warfarin Dosing. In *23rd USENIX Security Symposium (USENIX Security 14)*, pages 17–32, 2014. ISBN 978-1-931971-15-7.

Adrian Galdran, Gustavo Carneiro, and Miguel A. González Ballester. Balanced-MixUp for Highly Imbalanced Medical Image Classification. In Marleen de Bruijne, Philippe C. Cattin, Stéphane Cotin, Nicolas Padoy, Stefanie Speidel, Yefeng Zheng, and Caroline Essert, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pages 323–333, Cham, 2021. Springer International Publishing. ISBN 978-3-030-87240-3. doi: 10.1007/978-3-030-87240-3_31.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. ImageBind: One Embedding Space To Bind Them All. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190, 2023.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. MS-Celeb-1M: A Dataset and Benchmark for Large-Scale Face Recognition. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 87–102, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46487-9. doi: 10.1007/978-3-319-46487-9_6.

Omid Hajihassnai, Omid Ardakanian, and Hamzeh Khazaei. ObscureNet: Learning Attribute-invariant Latent Representation for Anonymizing Sensor Data. In *Proceedings of the International Conference on Internet-of-Things Design and Implementation*, pages 40–52, Charlottesvle VA USA, May 2021. ACM. ISBN 978-1-4503-8354-7. doi: 10.1145/3450268.3453534.

Haibo He and Edwardo A. Garcia. Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, September 2009. ISSN 1558-2191. doi: 10.1109/TKDE.2008.239.

Haibo He, Yang Bai, Edwardo A. Garcia, and Shutao Li. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, pages 1322–1328, June 2008. doi: 10.1109/IJCNN.2008.4633969.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the Knowledge in a Neural Network, March 2015.

Håkon Hukkelås, Rudolf Mester, and Frank Lindseth. DeepPrivacy: A Generative Adversarial Network for Face Anonymization. In George Bebis, Richard Boyle, Bahram Parvin, Darko Koracin, Daniela Ushizima, Sek Chai, Shinjiro Sueda, Xin Lin, Aidong Lu, Daniel Thalmann, Chaoli Wang, and Panpan Xu, editors, *Advances in Visual Computing*, pages 565–578, Cham, 2019. Springer International Publishing. ISBN 978-3-030-33720-9. doi: 10.1007/978-3-030-33720-9_44.

Muhammad Abdullah Jamal, Matthew Brown, Ming-Hsuan Yang, Liqiang Wang, and Boqing Gong. Rethinking Class-Balanced Methods for Long-Tailed Visual Recognition From a Domain Adaptation Perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7610–7619, 2020.

James Jordon, Jinsung Yoon, and Mihaela van der Schaar. PATE-GAN: Generating Synthetic Data with Differential Privacy Guarantees. In *International Conference on Learning Representations*, September 2018.

Lie Ju, Xin Wang, Lin Wang, Tongliang Liu, Xin Zhao, Tom Drummond, Dwarikanath Mahapatra, and Zongyuan Ge. Relational Subsets Knowledge Distillation for Long-Tailed Retinal Diseases Recognition. In Marleen de Bruijne, Philippe C. Cattin, Stéphane Cotin, Nicolas Padoy, Stefanie Speidel, Yefeng Zheng, and Caroline Essert, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pages 3–12, Cham, 2021. Springer International Publishing. ISBN 978-3-030-87237-3. doi: 10.1007/978-3-030-87237-3_1.

Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, Rafael G. L. D'Oliveira, Hubert Eichner, Salim El Rouayheb, David Evans, Josh Gardner, Zachary Garrett, Adrià Gascón, Badih Ghazi, Phillip B. Gibbons, Marco Gruteser, Zaid Harchaoui, Chaoyang He, Lie He, Zhouyuan Huo, Ben Hutchinson, Justin Hsu, Martin Jaggi, Tara Javidi, Gauri Joshi, Mikhail Khodak, Jakub Konecný, Aleksandra Korolova, Farinaz Koushanfar, Sanmi Koyejo, Tancrède Lepoint, Yang Liu, Prateek Mittal, Mehryar Mohri, Richard Nock, Ayfer Özgür, Rasmus Pagh, Hang Qi, Daniel Ramage, Ramesh Raskar, Mariana Raykova, Dawn Song, Weikang Song, Sebastian U. Stich, Ziteng Sun, Ananda Theertha Suresh, Florian Tramèr, Praneeth Vepakomma, Jianyu Wang, Li Xiong, Zheng Xu, Qiang Yang, Felix X. Yu, Han Yu, and Sen Zhao. Advances and Open Problems in Federated Learning. *Foundations and Trends® in Machine Learning*, 14(1–2): 1–210, June 2021. ISSN 1935-8237, 1935-8245. doi: 10.1561/2200000083.

Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling Representation and Classifier for Long-Tailed Recognition. In *International Conference on Learning Representations*, September 2019.

Bingyi Kang, Yu Li, Sa Xie, Zehuan Yuan, and Jiashi Feng. Exploring Balanced Feature Spaces for Representation Learning. In *International Conference on Learning Representations*, October 2020.

Jaehyung Kim, Jongheon Jeong, and Jinwoo Shin. M2m: Imbalanced Classification via Major-to-Minor Translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13896–13905, 2020.

Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes, 2013.

Alex Krizhevsky and Geoffrey Hinton. Learning Multiple Layers of Features from Tiny Images. Technical report, 2009.

Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, November 1998. ISSN 1558-2256. doi: 10.1109/5.726791.

Joon-Woo Lee, Hyungchul Kang, Yongwoo Lee, Woosuk Choi, Jieun Eom, Maxim Deryabin, Eunsang Lee, Junghyun Lee, Donghoon Yoo, Young-Sik Kim, and Jong-Seon No. Privacy-Preserving Machine Learning With Fully Homomorphic Encryption for Deep Neural Network. *IEEE Access*, 10:30039–30054, 2022. ISSN 2169-3536. doi: 10.1109/ACCESS.2022.3159694.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *Proceedings of the 39th International Conference on Machine Learning*, pages 12888–12900. PMLR, June 2022.

Shuang Li, Kaixiong Gong, Chi Harold Liu, Yulin Wang, Feng Qiao, and Xinjing Cheng. MetaSAug: Meta Semantic Augmentation for Long-Tailed Visual Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5212–5221, 2021.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft COCO: Common Objects in Context, February 2015.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal Loss for Dense Object Detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2980–2988, 2017.

Jialun Liu, Yifan Sun, Chuchu Han, Zhaopeng Dou, and Wenhui Li. Deep Representation Learning on Long-Tailed Data: A Learnable Embedding Augmentation Perspective. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2967–2976, Seattle, WA, USA, June 2020. IEEE. ISBN 978-1-72817-168-5. doi: 10.1109/CVPR42600.2020.00304.

Jiashuo Liu, Zheyan Shen, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, and Peng Cui. Towards Out-Of-Distribution Generalization: A Survey. https://arxiv.org/abs/2108.13624v2, August 2021.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach, July 2019a.

Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X. Yu. Large-Scale Long-Tailed Recognition in an Open World. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2532–2541, Long Beach, CA, USA, June 2019b. IEEE. ISBN 978-1-72813-293-8. doi: 10.1109/CVPR.2019.00264.

Gengchen Mai, Chris Cundy, Kristy Choi, Yingjie Hu, Ni Lao, and Stefano Ermon. Towards a foundation model for geospatial artificial intelligence (vision paper). In *Proceedings of the 30th International Conference on Advances in Geographic Information Systems*, SIGSPATIAL '22, pages 1–4, New York, NY, USA, November 2022. Association for Computing Machinery. ISBN 978-1-4503-9529-8. doi: 10.1145/3557915.3561043.

Mohammad Malekzadeh, Richard G. Clegg, and Hamed Haddadi. Replacement AutoEncoder: A Privacy-Preserving Algorithm for Sensory Data Analysis. In *2018 IEEE/ACM Third International Conference on Internet-of-Things Design*

and Implementation (IoTDI), pages 165–176, April 2018. doi: 10.1109/IoTDI.2018.00025.

Mohammad Malekzadeh, Richard G. Clegg, Andrea Cavallaro, and Hamed Haddadi. Mobile sensor data anonymization. In *Proceedings of the International Conference on Internet of Things Design and Implementation*, pages 49–58, Montreal Quebec Canada, April 2019. ACM. ISBN 978-1-4503-6283-2. doi: 10.1145/3302505.3310068.

Blaž Meden, Žiga Emeršič, Vitomir Štruc, and Peter Peer. K-Same-Net: K-Anonymity with Generative Deep Neural Networks for Face Deidentification. *Entropy*, 20(1):60, January 2018. ISSN 1099-4300. doi: 10.3390/e20010060.

Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M. Krumholz, Jure Leskovec, Eric J. Topol, and Pranav Rajpurkar. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265, April 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-05881-4.

E.M. Newton, L. Sweeney, and B. Malin. Preserving privacy by de-identifying face images. *IEEE Transactions on Knowledge and Data Engineering*, 17(2):232–243, February 2005. ISSN 1041-4347. doi: 10.1109/TKDE.2005.32.

Tung Nguyen, Johannes Brandstetter, Ashish Kapoor, Jayesh K. Gupta, and Aditya Grover. ClimaX: A foundation model for weather and climate, December 2023.

Paraskevi Nousi, Sotirios Papadopoulos, Anastasios Tefas, and Ioannis Pitas. Deep autoencoders for attribute preserving face de-identification. *Signal Processing: Image Communication*, 81:115699, February 2020. ISSN 09235965. doi: 10.1016/j.image.2019.115699.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning Robust Visual Features without Supervision, April 2023.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Sayak Paul and Pin-Yu Chen. Vision Transformers Are Robust Learners. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(2):2071–2081, June 2022. ISSN 2374-3468, 2159-5399. doi: 10.1609/aaai.v36i2.20103.

Steven T. Piantadosi. Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic bulletin & review*, 21(5):1112–1130, October 2014. ISSN 1069-9384. doi: 10.3758/s13423-014-0585-6.

Alec Radford, Jeff Wu, R. Child, D. Luan, Dario Amodei, and I. Sutskever. Language Models are Unsupervised Multitask Learners. *OpenAI blog*, 1(8):9, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, July 2021.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer, October 2019.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical Text-Conditional Image Generation with CLIP Latents, April 2022.

Jiawei Ren, Cunjun Yu, shunan sheng, Xiao Ma, Haiyu Zhao, Shuai Yi, and hongsheng Li. Balanced Meta-Softmax for Long-Tailed Visual Recognition. In *Advances in Neural Information Processing Systems*, volume 33, pages 4175–4186. Curran Associates, Inc., 2020.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models, April 2022.

Ruslan Salakhutdinov, Antonio Torralba, and Josh Tenenbaum. Learning to share visual appearance for multiclass object detection. In *CVPR 2011*, pages 1481–1488, June 2011. doi: 10.1109/CVPR.2011.5995720.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5B: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, December 2022.

Eli Schwartz, Leonid Karlinsky, Joseph Shtok, Sivan Harary, Mattias Marder, Abhishek Kumar, Rogerio Feris, Raja Giryes, and Alex Bronstein. Delta-encoder: An effective sample synthesis method for few-shot object recognition. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Amna Shahid and Malaika Mushtaq. A Survey Comparing Specialized Hardware And Evolution In TPUs For Neural Networks. In *2020 IEEE 23rd International*

Multitopic Conference (INMIC), pages 1–6, Bahawalpur, Pakistan, November 2020. IEEE. ISBN 978-1-72819-893-4. doi: 10.1109/INMIC50486.2020.9318136.

Ali Shahin Shamsabadi, Brij Mohan Lal Srivastava, Aurélien Bellet, Nathalie Vauquier, Emmanuel Vincent, Mohamed Maouche, Marc Tommasi, and Nicolas Papernot. Differentially Private Speaker Anonymization, October 2022.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership Inference Attacks Against Machine Learning Models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18, May 2017. doi: 10.1109/SP.2017.41.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning Structured Output Representation using Deep Conditional Generative Models. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

Latanya Sweeney. K-ANONYMITY: A MODEL FOR PROTECTING PRIVACY. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570, October 2002. ISSN 0218-4885, 1793-6411. doi: 10.1142/S0218488502001648.

Jingru Tan, Changbao Wang, Buyu Li, Quanquan Li, Wanli Ouyang, Changqing Yin, and Junjie Yan. Equalization Loss for Long-Tailed Object Recognition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11659–11668, Seattle, WA, USA, June 2020. IEEE. ISBN 978-1-72817-168-5. doi: 10.1109/CVPR42600.2020.01168.

Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, Anil Palepu, Basil Mustafa, Aakanksha Chowdhery, Yun Liu, Simon Kornblith, David Fleet, Philip Mansfield, Sushant Prakash, Renee Wong, Sunny Virmani, Christopher Semturs, S. Sara Mahdavi, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, Joelle Barral, Dale Webster, Greg S. Corrado, Yossi Matias, Karan Singhal, Pete Florence, Alan Karthikesalingam, and Vivek Natarajan. Towards Generalist Biomedical AI. *NEJM AI*, 1(3):AIoa2300138, February 2024. doi: 10.1056/AIoa2300138.

Grant Van Horn and Pietro Perona. The Devil is in the Tails: Fine-grained Classification in the Wild, September 2017.

Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The iNaturalist Species Classification and Detection Dataset, April 2018.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. Van Der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul Van Mulbregt, SciPy 1.0 Contributors, Aditya Vijaykumar, Alessandro Pietro Bardelli, Alex Rothberg, Andreas Hilboll, Andreas Kloeckner, Anthony Scopatz, Antony Lee, Ariel Rokem, C. Nathan Woods, Chad Fulton, Charles Masson, Christian Häggström, Clark Fitzgerald, David A. Nicholson, David R. Hagen, Dmitrii V. Pasechnik, Emanuele Olivetti, Eric Martin, Eric Wieser, Fabrice Silva, Felix Lenders, Florian Wilhelm, G. Young, Gavin A. Price, Gert-Ludwig Ingold, Gregory E. Allen, Gregory R. Lee, Hervé Audren, Irvin Probst, Jörg P. Dietrich, Jacob Silterra, James T Webber, Janko Slavič, Joel Nothman, Johannes Buchner, Johannes Kulick, Johannes L. Schönberger, José Vinícius De Miranda Cardoso, Joscha Reimer, Joseph Harrington, Juan Luis Cano Rodríguez, Juan Nunez-Iglesias, Justin Kuczynski, Kevin Tritz, Martin Thoma, Matthew Newville, Matthias Kümmerer, Maximilian Bolingbroke, Michael Tartre, Mikhail Pak, Nathaniel J. Smith, Nikolai Nowaczyk, Nikolay Shebanov, Oleksandr Pavlyk, Per A. Brodtkorb, Perry Lee, Robert T. McGibbon, Roman Feldbauer, Sam Lewis, Sam Tygier, Scott Sievert, Sebastiano Vigna, Stefan Peterson, Surhud More, Tadeusz Pudlik, Takuya Oshima, Thomas J. Pingel, Thomas P. Robitaille, Thomas Spura, Thouis R. Jones, Tim Cera, Tim Leslie, Tiziano Zito, Tom Krauss, Utkarsh Upadhyay, Yaroslav O. Halchenko, and Yoshiki Vázquez-Baeza. SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, 17(3):261–272, March 2020. ISSN 1548-7091, 1548-7105. doi: 10.1038/s41592-019-0686-2.

Jianfeng Wang, Thomas Lukasiewicz, Xiaolin Hu, Jianfei Cai, and Zhenghua Xu. RSG: A Simple but Effective Module for Learning Imbalanced Datasets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3784–3793, 2021.

Shuo Wang and Xin Yao. Multiclass Imbalance Problems: Analysis and Potential Solutions. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 42(4):1119–1130, August 2012. ISSN 1941-0492. doi: 10.1109/TSMCB.2012.2187280.

Yaqing Wang, Quanming Yao, James T. Kwok, and Lionel M. Ni. Generalizing from a Few Examples: A Survey on Few-shot Learning. *ACM Computing Surveys*, 53 (3):63:1–63:34, June 2020. ISSN 0360-0300. doi: 10.1145/3386252.

Benjamin Weggenmann, Valentin Rublack, Michael Andrejczuk, Justus Mattern, and Florian Kerschbaum. DP-VAE: Human-Readable Text Anonymization for Online Reviews with Differentially Private Variational Autoencoders. In *Proceedings of the ACM Web Conference 2022*, WWW '22, pages 721–731, New York, NY, USA,

April 2022. Association for Computing Machinery. ISBN 978-1-4503-9096-5. doi: 10.1145/3485447.3512232.

Tong Wu, Qingqiu Huang, Ziwei Liu, Yu Wang, and Dahua Lin. Distribution-Balanced Loss for Multi-label Classification in Long-Tailed Datasets. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 162–178, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58548-8. doi: 10.1007/978-3-030-58548-8_10.

Yifan Wu, Fan Yang, Yong Xu, and Haibin Ling. Privacy-Protective-GAN for Privacy Preserving Face De-Identification. *Journal of Computer Science and Technology*, 34(1):47–60, January 2019. ISSN 1000-9000, 1860-4749. doi: 10.1007/s11390-019-1898-8.

Liyang Xie, Kaixiang Lin, Shu Wang, Fei Wang, and Jiayu Zhou. Differentially Private Generative Adversarial Network, February 2018.

Zuobin Xiong, Wei Li, Qilong Han, and Zhipeng Cai. Privacy-Preserving Auto-Driving: A GAN-Based Approach to Protect Vehicular Camera Data. In *2019 IEEE International Conference on Data Mining (ICDM)*, pages 668–677, Beijing, China, November 2019. IEEE. ISBN 978-1-72814-604-1. doi: 10.1109/ICDM.2019.00077.

Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. MedMNIST v2 - A large-scale lightweight benchmark for 2D and 3D biomedical image classification. *Scientific Data*, 10(1):41, January 2023. ISSN 2052-4463. doi: 10.1038/s41597-022-01721-8.

Xi Yin, Xiang Yu, Kihyuk Sohn, Xiaoming Liu, and Manmohan Chandraker. Feature Transfer Learning for Face Recognition With Under-Represented Data. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5697–5706, Long Beach, CA, USA, June 2019. IEEE. ISBN 978-1-72813-293-8. doi: 10.1109/CVPR.2019.00585.

Yuhang Zang, Chen Huang, and Chen Change Loy. FASA: Feature Augmentation and Sampling Adaptation for Long-Tailed Instance Segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3457–3466, 2021.

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. Mixup: Beyond Empirical Risk Minimization, April 2018.

Xiao Zhang, Zhiyuan Fang, Yandong Wen, Zhifeng Li, and Yu Qiao. Range Loss for Deep Face Recognition with Long-Tailed Training Data. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5419–5428, Venice, October 2017. IEEE. ISBN 978-1-5386-1032-9. doi: 10.1109/ICCV.2017.578.

Yifan Zhang, Bryan Hooi, Lanqing Hong, and Jiashi Feng. Self-Supervised Aggregation of Diverse Experts for Test-Agnostic Long-Tailed Recognition.

Advances in Neural Information Processing Systems, 35:34077–34090, December 2022a.

Yifan Zhang, Bingyi Kang, Bryan Hooi, Shuicheng Yan, and Jiashi Feng. Deep Long-Tailed Learning: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10795–10816, September 2023. ISSN 0162-8828, 2160-9292, 1939-3539. doi: 10.1109/TPAMI.2023.3268118.

Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D. Manning, and Curtis P. Langlotz. Contrastive Learning of Medical Visual Representations from Paired Images and Text. In *Proceedings of the 7th Machine Learning for Healthcare Conference*, pages 2–25. PMLR, December 2022b.

Zhisheng Zhong, Jiequan Cui, Shu Liu, and Jiaya Jia. Improving Calibration for Long-Tailed Recognition. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16484–16493, Nashville, TN, USA, June 2021. IEEE. ISBN 978-1-66544-509-2. doi: 10.1109/CVPR46437.2021.01622.

Boyan Zhou, Quan Cui, Xiu-Shen Wei, and Zhao-Min Chen. BBN: Bilateral-Branch Network With Cumulative Learning for Long-Tailed Visual Recognition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9716–9725, Seattle, WA, USA, June 2020. IEEE. ISBN 978-1-72817-168-5. doi: 10.1109/CVPR42600.2020.00974.

Yukun Zhou, Mark A. Chia, Siegfried K. Wagner, Murat S. Ayhan, Dominic J. Williamson, Robbert R. Struyven, Timing Liu, Moucheng Xu, Mateo G. Lozano, Peter Woodward-Court, Yuka Kihara, Andre Altmann, Aaron Y. Lee, Eric J. Topol, Alastair K. Denniston, Daniel C. Alexander, and Pearse A. Keane. A foundation model for generalizable disease detection from retinal images. *Nature*, 622(7981): 156–163, October 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06555-x.

Xiangxin Zhu, Dragomir Anguelov, and Deva Ramanan. Capturing Long-tail Distributions of Object Subcategories. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 915–922, 2014.

Alexander Ziller, Dmitrii Usynin, Rickmer Braren, Marcus Makowski, Daniel Rueckert, and Georgios Kaissis. Medical imaging deep learning with differential privacy. *Scientific Reports*, 11(1):13524, June 2021. ISSN 2045-2322. doi: 10.1038/s41598-021-93030-0.

G. K. Zipf. *The Psycho-Biology of Language*. The Psycho-Biology of Language. Houghton, Mifflin, Oxford, England, 1935.

Declaration of Authorship

Ich erkläre hiermit gemäß §9 Abs. 12 APO, dass ich die vorstehende Abschlussarbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Des Weiteren erkläre ich, dass die digitale Fassung der gedruckten Ausfertigung der Abschlussarbeit ausnahmslos in Inhalt und Wortlaut entspricht und zur Kenntnis genommen wurde, dass diese digitale Fassung einer durch Software unterstützten, anonymisierten Prüfung auf Plagiate unterzogen werden kann.

Place, Date	Signature